Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

A Many-outcomes Perspective on the Synthetic Control Method

Yixin Wang*
YIXINW@UMICH.EDU
Department of Statistics
University of Michigan

Aaron Schein*
SCHEIN@UCHICAGO.EDU
Department of Statistics
University of Chicago

Joshua Shou
JSHOU@UCHICAGO.EDU
Department of Statistics
University of Chicago

David M. Blei
DAVID.BLEI@COLUMBIA.EDU
Department of Statistics
Department of Computer Science
Columbia University

Editor: Kevin Murphy and Bernhard Schölkopf

Abstract

The synthetic controls method is a popular approach for analyzing panel data to estimate the causal effect of major policy interventions (Abadie et al., 2010, 2015; Abadie & Gardeazabal, 2003). The method is known to be justified when the data-generating process can be described by a linear factor model (Abadie et al., 2010). This paper provides an alternative justification for the synthetic control method, which leads to a generalization of the algorithm that can incorporate non-linear factor models. Our justification centers around *negative control outcomes* (Lipsitch et al., 2010; Rosenbaum, 1989), outcome variables for which the intervention is known *a priori* to have no effect. In synthetic controls, both pre-intervention outcomes of the target and pre- and post-intervention outcomes of the donors are all negative controls. Our key observation is that causal effects are identifiable under two assumptions: (1) the latent factors of the true (not necessarily linear) factor model satisfy weak unconfoundedness, and (2) the latent factors can be theoretically determined by the negative control outcomes. For example, these assumptions can be approximately satisfied in panel data with a long pre-treatment period. Based on these assumptions, we then generalize the synthetic controls method to an algorithm that estimates causal effects in non-linear settings. Crucial to this procedure is the idea of model criticism: the practitioner has license to explore and select a probabilistic factor model from a large class, provided it passes its goodness-of-fit test. We demonstrate this approach on a semi-synthetic study and use it to re-analyze a recent study on the impact of opening football stadiums during the COVID-19 pandemic (García Bulle et al., 2022). We find that this generalized approach to synthetic controls provides better counterfactual prediction than classical synthetic controls.

Keywords: Causal inference, synthetic control, probabilistic models

1. Introduction

Researchers often analyze *panel data* to assess the effects of major policy changes, like the implementation of new state laws, or the enforcement of new public health measures. A panel data set comprises repeated observations of units over time, and thus provides a "before-and-after" view of changes in policy that happen during the observation window. Panel data can be particularly useful for evaluating policies when the outcomes of certain units are *negative controls* (Lipsitch et al., 2010; Rosenbaum, 1989)—i.e., they are known a priori to be unaffected by the policy change of interest. More specifically, negative control outcomes can provide information about the *counterfactual* world in which the policy change did not occur. In this paper, we seek to re-interpret and generalize the method of synthetic controls (SC) (Abadie et al., 2010, 2015; Abadie & Gardeazabal, 2003)—a popular approach for estimating counterfactual outcomes from panel data—through the lens of negative control outcomes.

To ground our discussion, consider the following recent application of synthetic controls by García Bulle et al. (2022) which we will use as a running example throughout the paper. During September 2020, at the height of the COVID-19 pandemic, the National Football League (NFL) decided to allow its stadiums to reopen to fans. What impact did this policy change have on the spread of COVID-19? To address this question, García Bulle et al. collected daily panel data of COVID-19 case counts in US counties containing NFL stadiums. Their approach then hinged upon the key fact that the NFL's decision only affected stadiums in counties whose COVID-19 protocols permitted large outdoor public gatherings. Thus, while Dallas Cowboys fans returned to their stadium, New York Giants fans continued to watch from home. We will say Dallas County is a *treated* unit since its protocols allowed Cowboys fans to return, while Bergen County, where the Giants play, is a *control* unit.

Applying the synthetic control method to this setting involves two high-level steps. First, the method finds a weighted combination of the control counties' outcomes that best reconstructs the treated units' outcomes during the pre-intervention period, prior to the NFL's decision. Second, the method then uses that learned weighting to construct a counterfactual estimate of the treated counties' outcomes in the post-intervention period, after the NFL's decision. Intuitively, if the weighted combination accurately captures the relationship between control and treated units before the intervention, the same weights may then produce valid predictions of the COVID-19 case counts in the treated counties for the counterfactual world in which they did not reopen their stadiums. In following these steps, García Bulle et al. find that the NFL's decision to reopen likely had no impact on the spread of COVID-19.

What justifies the synthetic control method? Specifically, when does an estimated weighted combination of control outcomes provide an unbiased estimate of the counterfactual for treated ones? The condition most commonly cited is when the matrix of pre-intervention outcomes is low rank, such that any entry can be written as coming from a *linear factor model*. In such a setting, a weighted combination emerges as a natural consequence of the factor structure, allowing us to express one unit's outcome as a weighted combination of others'.

This paper rethinks the underlying assumptions that justify the synthetic control method. In particular, we show that the method can be justified without appealing to linearity assumptions. Our alternative justification centers around *negative control outcomes* (Lipsitch et al., 2010; Rosenbaum, 1989), outcome variables for which the intervention is known *a priori* to have no effect. The presence of such variables is what characterizes the typical synthetic control setting, wherein *all* observations are negative controls, except those for treated units in the post-intervention period. In other words, the pre-treatment outcomes serve as the negative control outcomes here since a treatment cannot

affect the outcome variables that precede it. More specifically, we show that the average treatment effect on the treated (ATT) is identifiable under two assumptions:

1. the factors of a (not necessarily linear) factor model for the untreated outcomes satisfy weak unconfoundedness, and
2. the latent factors can be determined in theory[1] by some negative control outcomes.

These assumptions can be approximately satisfied in panel data with a long pre-treatment period and many control units. Given these assumption, we develop a generalization of the synthetic control method using probabilistic factor models, namely an algorithm that estimates causal effects in the presence of many negative control outcomes. It first fits a (possibly non-linear) probabilistic factor model (e.g. probabilistic principal component analysis (Tipping & Bishop, 1999) or a mixture model (McLachlan & Basford, 1988)) to all the outcomes, both control and non-control. We term the fitted latent factors as the *estimated confounding structure*; they may not coincide exactly with the true unobserved confounders, but, under suitable conditions, they consistently estimate some invertible functions of the true unobserved confounders. It then uses the estimated confounding structure in a downstream causal inference. We prove that, under suitable assumptions, the estimated confounding structure can correct for *multi-period confounders*, those that affect the treatment, the study outcome, and the control outcomes. And we show how the seminal method of synthetic controls (Abadie et al., 2010, 2015; Abadie & Gardeazabal, 2003) can be viewed as a special case of our algorithm, expanding the settings where synthetic controls can be used.

In practice, a key component in applying our approach is the search for a probabilistic factor model that fits the data, which often involves an iterative cycle of model-building, -fitting, and -critiquing (Blei, 2014). We illustrate this in a case study where we consider multiple candidate models until we find one that fits according to a holdout predictive check (Moran et al., 2024). In general, this perspective opens up a range of model-based approaches tailored to panel data of counts, categorical, or ordinal outcomes, among forms of non-standard or complex observations. Indeed, in semi-synthetic and real empirical studies, we find that the proposed algorithm can enable better counterfactual prediction via its use of non-Gaussian factor models.

While these assumptions guarantee the validity of the proposed algorithm, they may not hold or only approximately hold in practice. In particular, the "determinable in theory" assumption can only hold in the limit of observing an infinite number of control outcomes. Given a finite number of control outcomes in practice, it is only possible to approximately determine the latent factors. Further, whether the estimated confounding structure can correct for multi-outcome confounders in practice relies on the well-specification of the factor model. The estimated confounding structure produced by misspecified factor models can lead to biased downstream causal inferences.

Related work. This work build on multiple threads of research in causal inference.

Causal inference on panel data. Causal inference on panel data has been studied from many perspectives. Brodersen et al. (2015) approaches it with Bayesian structural time-series models. Bertrand et al. (2004); Card (1990) studies it with differences-in-difference estimates, along with more recent works of Goodman-Bacon (2018); Callaway & Sant'Anna (2019); Abraham & Sun (2018); Imai & Kim (2019); Cengiz et al. (2019); Heckman et al. (1998, 1997); Abadie (2005); Athey & Imbens (2006); Qin et al. (2008); Bonhomme & Sauder (2011); Botosaru & Gutierrez (2018); Bonhomme & Sauder (2011); De Chaisemartin & D'HaultfŒuille (2017); Callaway et al. (2018).

1. We will define "determined in theory" rigorously later. Loosely, it means the negative control outcomes can tell us precisely what the values of the latent factors are.

Abadie et al. (2010, 2015); Abadie & Gardeazabal (2003) approach causal estimation on panel data with the synthetic control method, along with many variants, extensions, and discussions (Robbins et al., 2017; Doudchenko & Imbens, 2016; Amjad et al., 2019, 2018; O'Neill et al., 2016; Ferman, 2019; Xu, 2017; Dusetzina et al., 2015; Gobillon & Magnac, 2016; Li, 2019; Kinn, 2018; Chernozhukov et al., 2017; Ben-Michael et al., 2022; Chernozhukov et al., 2018; Abadie & L'Hour, 2017; Hsiao et al., 2012; Samartsidis et al., 2019; Viviano & Bradic, 2019; Dube & Zipperer, 2015; Gunsilius, 2023; Chen, 2023; Agarwal et al., 2023; Tanaka, 2021; Spiess et al., 2024; Bottmer et al., 2024). Taking the matrix completion perspective, Athey et al. (2018) unifies the synthetic control approaches with the unconfoundedness approaches (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015). Arkhangelsky et al. (2019) develops synthetic difference-in-differences, which generalizes the synthetic control method and the difference-in-difference estimator. Many further developments succeed; examples include Ben-Michael et al. (2021) that extends to settings where perfect pre-treatment fit is infeasible, Ben-Michael et al. (2022) that extends the synthetic control method to staggered adoption, Sun et al. (2023) that proposes to use multiple outcomes to improve synthetic control, and many others.

This work studies the more general setting of multiple control outcomes than the setting of panel data. The identification strategy we develop do not rely on linearity of the structural models as in generalizations of the synthetic control method (Abadie et al., 2010; Xu, 2017) and further generalizations to matrix completion settings (Amjad et al., 2018, 2019; Agarwal et al., 2023). This work also differs from other existing alternative identification conditions for the synthetic control method like O'Neill et al. (2016); Kinn (2018); Zeitler et al. (2023); Shi et al. (2022) in drawing on the asymptotics of latent variable models (Chen et al., 2017). For example, Shi et al. (2022) relies on an invariant-causal-mechanism assumption for the latent confounder, which is closely related to our (nonlinear) factor-model specification. However, it restricts attention to discrete latent confounders and achieves identification via imposing a cardinality constraint, specifically, bounding the number of donors used in weighting relative to the support size of the confounder, a set of constraints our framework does not require.

Negative controls. Negative controls, also known as proxy variables, are often used to detect the presence of unobserved confounding (Lipsitch et al., 2010; Rosenbaum, 1989). More recently, researchers have developed strategies that use negative controls to identify causal effects under unobserved confounding; they posit different assumptions under which certain observed or counterfactual variables suffice to account for unobserved confounding (Tchetgen Tchetgen, 2013; Miao & Tchetgen, 2018; Egami, 2018; Kasza et al., 2017; Sanderson et al., 2017; Shi et al., 2018; Miao & Tchetgen Tchetgen, 2017; Flanders et al., 2017; Richardson et al., 2015; Sofer et al., 2016; Arnold et al., 2016; Dusetzina et al., 2015). More recently, Miao et al. (2018); Kuroki & Pearl (2014) leverage completeness conditions on the distributions of variables to identify causal effects with negative controls. Schuemie et al. (2016); Madigan et al. (2014); Schuemie et al. (2018) use negative controls to calibrate p-values of average treatment effects in observational studies. Zheng et al. (2023) and Zhou et al. (2024) extend related ideas to synthetic control and multi-outcome settings with factor-structured outcomes, without requiring access to pre-treatment outcomes. In particular, Zheng et al. (2023) develop a sensitivity analysis framework for unobserved confounders that are shared across multiple outcomes, while Zhou et al. (2024) establish causal identification for factor-structured outcomes by extending the proxy variable identification strategy of Miao et al. (2018). Both works focus on a finite number of factor-structured outcomes and do not rely on pre-treatment data. In contrast, our work considers a single outcome of interest, leverages factor-structured pre-treatment outcomes, and achieves causal identification through the consistent estimation of latent factor models as the number of outcomes tends to infinity.

Finally, our proposed algorithm is related to some of the ideas and techniques from multi-treatment causal inference (Wang & Blei, 2019a, 2020) that leverages many treatments as evidence about unobserved confounders. However, the proposed algorithm operates in a different setting with a single treatment, a study outcome, and many control outcomes; it leverages many control outcomes as evidence about unobserved multi-period confounders. There has been scholarly debate about some of the identification theory behind Wang & Blei (2019a), particularly Ogburn et al. (2019, 2020) and Wang & Blei (2020, 2019b). Much of the debate surrounds the valid use of multiple treatments for causal identification, which does not pertain to this paper.

Unconfoundedness and sensitivity analysis. Rosenbaum & Rubin (1983); Hirano & Imbens (2004); Imbens (2000) show that weak unconfoundedness is a key assumption that help identify causal effects. Many works attempt to assess the plausibility of these assumptions with sensitivity analysis, including Franks et al. (2019); Robins et al. (2000); Gilbert et al. (2003); Imai & Van Dyk (2004); Imbens (2003). More recently, researchers have developed test for confounding leveraging external data (Sharma et al., 2016; Janzing & Schölkopf, 2018b,a; Liu & Chan, 2018). This work differs from them in that we leverage external data of multiple control outcomes to adjust for unobserved confounding.

This paper. In Section 2, we detail the set up of the causal problem, discuss the linear factor model assumptions of the synthetic control method. Section 3 then proposes factor-model-like assumptions for causal identification in non-linear structural models. Under these assumptions, we propose an algorithm that estimates causal effects in non-linear settings, discuss the class of unobserved confounders it can handle, and describe the theoretical assumptions required for it to produce unbiased causal estimates. Finally, we discuss how the proposed algorithm and its surrounding theory generalizes the method of synthetic controls, focusing on settings with many treated units and a long pre-treatment period. Finally, we present a semi-synthetic study and a case study about COVID-19 in Section 5. We show that both the proposed algorithm and the synthetic control method can produce accurate causal estimates in the presence of unobserved confounding.

2. Panel data and the synthetic control method

Consider the following setting. We observe panel data, consisting of repeated measurements of n units' outcomes over time. We are interested in the impact of a major policy change or *intervention* that occurs at $t = 0$. Our panel consists of $\overleftarrow{m}$ time steps before the intervention, indexed by $t < 0$, and $\overrightarrow{m}$ time steps after, indexed by $t > 0$. We also know *a priori* that some units i were unaffected by the intervention, denoted by $A_i = 0$, while others were potentially affected, denoted by $A_i = 1$. Our goal is to estimate the causal effect of the intervention on the affected units' outcomes in the post-intervention period.

To think concretely about this setting, we return to the COVID-19 example in the introduction. Each unit $i \in [n]$ is a US county that contains an NFL stadium. The observed outcomes are the COVID-19 case counts $Y_{i,t} \in \mathbb{Z}_+$ in each county i at each time step t. The intervention at $t = 0$ is the NFL announcing in September 2020 that it would allow fans to return to stadiums. Some counties had public health protocols that permitted fans to return ($A_i = 1$), while other counties had stricter protocols, which did not ($A_i = 0$). We assume that the NFL's decision did not affect the spread of COVID-19 in counties where fans did not return to stadiums, and thus that the effect of the NFL's decision was entirely contained to those counties where fans did, in fact, return. Our goal is to estimate that effect.

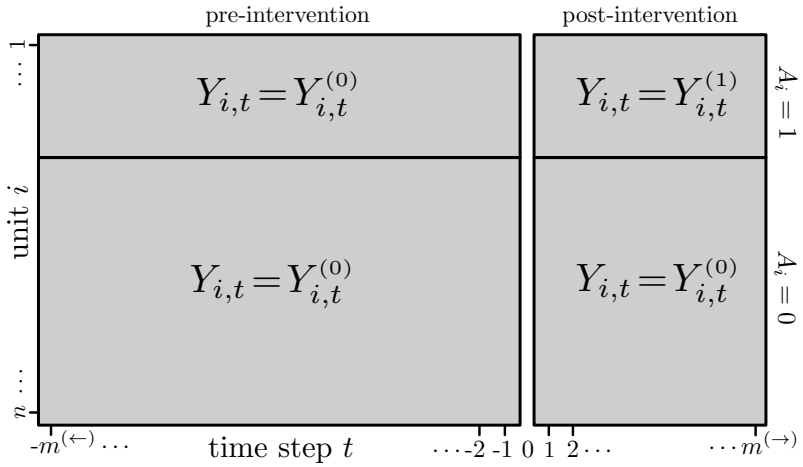


Figure 1: A panel data set in our setting is a (units × time)-matrix of observed outcomes $Y_{i,t}$. The observation window contains an intervention at $t = 0$. There are n units, some who are affected by the intervention ($A_i = 1$), and some who are not ($A_i = 0$). There are m time steps that are composed of $\overleftarrow{m}$ pre-intervention time steps ($t < 0$), and $\overrightarrow{m}$ post-intervention time steps ($t > 0$): $m = \overleftarrow{m} + \overrightarrow{m}$. For all pre-intervention time steps, and all control units, the observed outcomes coincide with the untreated potential outcomes—i.e., $Y_{i,t} = Y_{i,t}^{(0)}$ for $i < 0$ or $t < 0$. For only the treated units in post-intervention time steps, the observed outcomes coincide with treated potential outcomes—i.e., $Y_{i,t} = Y_{i,t}^{(1)}$ for $i > 0$ and $t > 0$.

What effect the NFL's decision had on affected counties is a causal question. Using the potential outcomes framework, let $Y_{i,t}^{(1)}$ denote the COVID-19 case count in county i during time step t in the possible world in which fans were attending games there and then. Let $Y_{i,t}^{(0)}$ denote the corresponding outcome in the world in which fans were not attending games. The observed outcomes are then equal to

$$Y_{i,t} = \begin{cases} Y_{i,t}^{(0)} & \text{if } t < 0 \quad \text{(pre-intervention)} \\ Y_{i,t}^{(A_i)} & \text{if } t > 0 \quad \text{(post-intervention)} \ . \end{cases} \tag{1}$$

A characteristic feature of this setting is that we observe *all* units' $Y_{i,t}^{(0)}$ before the intervention $t < 0$, while afterwards ($t > 0$) we observe some units' $Y_{i,t}^{(0)}$ and some units' $Y_{i,t}^{(1)}$ (see fig. 1).

By assumption, the causal effect of the NFL's intervention is contained to only the outcomes of the affected counties ($A_i = 1$) in the post-intervention period ($t > 0$). The effect is thus naturally defined as the average treatment effect on the treated (ATT):

$$\text{ATT} = \mathbb{E}\left[Y_{i,t}^{(1)} - Y_{i,t}^{(0)} \mid A_i = 1 \text{ and } t > 0\right]. \tag{2}$$

The ATT decomposes into the difference of two expectations, one factual, the other counterfactual. The expectation $\mathbb{E}\left[Y_{i,t}^{(1)} \mid A_i = 1 \text{ and } t > 0\right]$ is factual since it coincides with an observed quantity—i.e., it equals $\mathbb{E}\left[Y_{i,t} \mid A_i = 1 \text{ and } t > 0\right]$. By contrast, the expectation $\mathbb{E}\left[Y_{i,t}^{(0)} \mid A_i = 1 \text{ and } t > 0\right]$ is counterfactual since, in words, it represents what the post-intervention COVID-19 outcomes would have been in the affected counties had fans, counter to fact, not returned to games. The challenge in estimating the ATT is estimating this counterfactual expectation.

A popular approach to estimating this counterfactual is the method of synthetic control (SC) (Abadie et al., 2010, 2015; Abadie & Gardeazabal, 2003). At a high level, SC uses the post-intervention outcomes of the unaffected units—termed the "donors"—to predict the counterfactual outcomes of the affected units—termed the "targets." In more detail, for each target unit i, SC finds a weighted combination of the donors' pre-intervention outcomes that match the target's—i.e., it finds weights $\{w_{i,j}^{\text{SC}}\}_{j:\, A_j=0}$ that satisfy

$$\sum_{j:\, A_j=0} w_{i,j}^{\text{SC}} Y_{j,t} = Y_{i,t} \;\text{ for } t<0, \tag{3}$$

$$\text{s.t.} \sum_{j:A_j=0} w_{i,j}^{\text{SC}} = 1, \text{ and } w_{i,j}^{\text{SC}} \geq 0. \tag{4}$$

When covariates X_i are observed, the weights may further satisfy

$$\sum_{j:\, A_j=0} w_{i,j}^{\text{SC}} X_j = X_i. \tag{5}$$

With such weights, SC estimates the target unit's counterfactual outcomes with

$$\widehat{Y_{i,t}^{(0)}} \leftarrow \sum_{j:\, A_j=0} w_{i,j}^{\text{SC}}\, Y_{j,t} \;\text{ for } t>0. \tag{6}$$

The average treatment effect on the treated (ATT) is then estimated by averaging the difference between the factual and (estimated) counterfactual outcomes across targets:

$$\widehat{\text{ATT}} \leftarrow \frac{1}{\sum_{i=1}^{n} \mathbb{I}\{A_i=1\}} \sum_{i:\, A_i=1} \left(Y_{i,t} - \sum_{j:A_j=0} w_{i,j}^{\text{SC}}\, Y_{j,t} \right). \tag{7}$$

What justifies the synthetic control method? Specifically, under what conditions does Equation (7) provide an unbiased estimate of the ATT? The synthetic control has traditionally been justified by assuming that the observed panel data was generated by a linear structural model, i.e.,

$$Y_{i,t} = \lambda_t^\top X_i + \theta_t^\top U_i + \alpha\, A_i\, \mathbb{I}(t>0) + \epsilon_{i,t} \tag{8}$$

where $\left\{\theta_t, \lambda_t\right\}_t$ are parameters of the structural model, U_i are unobserved confounders, and $\left\{\epsilon_{i,t}\right\}_t$ are independent and identically distributed zero-mean random variables that satisfy $A_i, U_i, X_i \perp \left\{\epsilon_{i,t}\right\}_t$. Since $Y_{i,t} = Y_{i,t}^{\left(A_i \cdot \mathbb{I}(t>0)\right)}$, this implies the following structural model for potential outcomes:

$$\mathbb{E}\left[Y_{i,t}^{(0)}\right] = \lambda_t^\top X_i + \theta_t^\top U_i \tag{9}$$

$$\mathbb{E}\left[Y_{i,t}^{(1)}\right] = \mathbb{E}\left[Y_{i,t}^{(0)}\right] + \alpha\, \mathbb{I}(t>0) \tag{10}$$

where the expectation is taken with respect to the random noise variables $\epsilon_{i,t}$.

If the potential outcomes truly do follow the model in Equations (9) and (10), then the synthetic control method will provide an unbiased estimate of the ATT. At a high level, the reason is that the (residualized) untreated potential outcomes follow a linear latent factor model,

$$\mathbb{E}\left[Y_{i,t}^{(0)}\right] - \lambda_t^\top X_i = \theta_t^\top U_i. \tag{11}$$

The latent factors U_i for all units i, along with the factors θ_t for $t < 0$, can be consistently estimated from the pre-intervention outcomes under suitable conditions.[2] With U_i estimated, the factors θ_t for $t > 0$ can then be consistently estimated using the donors' post-intervention outcomes. Plugging in estimates of U_i and θ_t into Equation (8) then permits estimation of α, which coincides with the ATT.

While the assumption of a linear model enables identification, many practical settings do not satisfy it. For example, the outcomes in the COVID-19 running example are counts, which are unlikely to be generated by a linear factor model. How can we estimate causal effects when the data follows a potentially nonlinear factor model? In the next section, we show how the synthetic control method can be justified by an alternative set of assumptions that hold in settings beyond the special case of a linear structural model.

3. Generalizing the synthetic control method to nonlinear factor models

This section proposes an alternative set of assumptions that enable causal identification in the synthetic control setting. These assumptions are met by many structural models beyond linear ones and thus naturally lead to a generalization of the synthetic control method. We build on a probabilistic modeling interpretation of the synthetic control method given in the previous section to formulate these assumptions.

Our main idea is to replace the linearity assumption undergirding traditional synthetic control with the assumption that the underlying structural model is any (possibly non-linear) factor model that induces a certain kind of exchangeability in the outcomes. In particular, the unobserved confounder U_i is a random variable that rendered all of unit i's outcomes conditionally independent:

$$P\left(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i \,|\, U_i\right) = \left[\prod_{t=1}^{\overleftarrow{m}} P\left(Y_{i,-t} \,|\, U_i\right)\right] \left[\prod_{t=1}^{\overrightarrow{m}} P\left(Y_{i,t} \,|\, U_i\right)\right]. \tag{12}$$

As an example, the latent confounders U_i in the linear structural model (Equation (8)) satisfy this conditional independence.

While the outcomes $\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i$ satisfy a factor model given U_i, it does not imply that one can fit a factor model and take its estimate of U_i to perform the confounding adjustment. However, we will show that, under certain conditions, such a procedure is legitimate. This argument will license us to use estimates of U_i—which we denote as Z_i—in place of U_i for confounder adjustment. (We keep the estimate Z_i as a separate notation as U_i since factor models may not be identifiable and the estimate Z_i may only be consistent with U_i up to some bijective transformation.)

The conditions under which Z_i is a valid estimate for U_i in confounder adjustment are two-fold: (1) U_i is an *emancipator* that can be determined in theory, as defined next in Section 3.1, and (2) U_i is a *multi-period confounder*, as defined in Section 3.2. In Section 3.3 we detail our strategy for identifying causal effects when these conditions are met, and in **??** we show how the method of synthetic control can be viewed as a special case of this strategy.

Although we frame our discussion using panel data and pre-intervention outcomes (the typical context for the synthetic control method), the identification strategy itself is more general. It applies

2. Bai (2003) provides precise conditions (e.g. each factor has nontrivial contributions to the variance of $Y_{i,t}^{(0)}$) that guarantee the existence of consistent estimates for the factors θ_t.

to any setting where $\overleftarrow{\mathbf{Y}}_i$ represents "negative control outcomes", namely any variables known a priori to be unaffected by the treatment. Pre-treatment outcomes are a special case of negative controls. The strategy's validity does not depend on the time-series nature of panel data. For example, it could apply to a recommendation system where the treatment is showing users a Star Wars advertisement. The outcome $\overrightarrow{\mathbf{Y}}_i$ is whether they watch Star Wars, and $\overleftarrow{\mathbf{Y}}_i$ could be whether they watch unrelated content (e.g., a nature documentary). Here $\overleftarrow{\mathbf{Y}}_i$ is a negative control outcome that is unlikely to be affected by the Star Wars recommendation. Even without panel data, synthetic control applies if we perform adjustments based on these negative control outcomes; the identification strategy does not require time series structure. Nevertheless, we will stick to the pre-intervention outcomes and panel data framing here since synthetic control is most commonly presented in panel data contexts.

3.1 Emancipators that are determinable-in-theory

Consider a parametric model θ that describes the population distribution of unit i's outcomes as a marginal over latent variable U_i with prior distribution $P_\theta(U_i)$. Moreover, say the variable U_i, under this model, "emancipates" unit i's outcomes s—i.e., renders them conditionally independent:

$$P\left(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i\right) = \int \left[\prod_{t=1}^{\overleftarrow{m}} P_\theta\left(Y_{i,-t} \,|\, U_i\right)\right] \left[\prod_{t=1}^{\overrightarrow{m}} P_\theta\left(Y_{i,t} \,|\, U_i\right)\right] P_\theta(Z_i)\, \mathbf{d}U_i. \tag{13}$$

Suppose further that the variable U_i can be "determined in theory" by i's pre-intervention outcomes:

$$P_\theta\left(U_i \,|\, \overleftarrow{\mathbf{Y}}_i\right) = \delta\left(f\left(\overleftarrow{\mathbf{Y}}_i\right)\right), \tag{14}$$

where $\delta(x)$ represents a point mass at x, and f is some (unknown) deterministic function. "Determinable in theory" means we effectively observe U_i by observing pre-intervention outcomes. (These outcomes must precede the intervention, since we use them to infer and adjust for the unobserved confounder, so they cannot themselves be influenced by treatment. More generally, any negative control outcomes, i.e. variables known a priori to be unaffected by the treatment, can serve this role.) This property would hold under many models in the limit of infinitely many pre-intervention outcomes ($\overleftarrow{m} \to \infty$) were observed. For instance, it holds for the linear model that justifies the synthetic control method (Chen et al., 2017). We do not have access to infinitely many pre-intervention outcomes in practice, hence the name "determinable *in theory*."

A variable U_i with these two properties is what we will refer to as a *determinable-in-theory emancipator*:

Definition 1 (Determinable-in-theory emancipator)**.** *The unobserved confounder U_i is a determinable-in-theory emancipator under model θ if*

1. *it "emancipates"—i.e., renders conditionally independent—all of i's outcomes (Equation* (13)*),*
2. *it can be "determined in theory" by pre-intervention outcomes (Equation* (14)*).*

Existing literature provides a few examples of primitive conditions that will satisfy the determinable-in-theory condition on U. For example, Theorem 2 of Chen et al. (2020) states that if outcomes are

generated by a generalized linear factor model

$$P\left(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i \,|\, U_i\right) = \left[\prod_{t=1}^{\overleftarrow{m}} \exp\left(\frac{Y_{i,\text{-}t} m_{i,\text{-}t} - b(m_{i,\text{-}t})}{\phi} + c(Y_{i,\text{-}t}, \phi)\right)\right] \times \left[\prod_{t=1}^{\overrightarrow{m}} \exp\left(\frac{Y_{i,t} m_{it} - b(m_{it})}{\phi} + c(Y_{i,\text{-}t}, \phi)\right)\right],$$

where $m_{i,\text{-}t} = \theta_{\text{-}t}^\top U_i$, $m_{it} = \theta_t^\top U_i$, and $b(\cdot)$ and $c(\cdot)$ are prespecified functions that depend on the member of the exponential family. Then there exists a consistent estimator of U_i (up to shifting and scaling), under the assumption that (1) the limit

$$p_Q(S) \triangleq \lim_{\overleftarrow{m}\to\infty} \frac{|j : \theta_{jk} \neq 0, \text{ if } k \in S \text{ and } \theta_{jk} = 0, \text{ if } k \notin S,\, 1 \leq j \leq \overleftarrow{m}\}|}{\overleftarrow{m}}$$

exists for any subset $S \subset \{1, \ldots, K\}\}$, and $p_Q(\emptyset) = 0$, where K is the dimensionality of θ_j and θ_{jk} is the k-dimension of θ_j, (2) the natural parameter space $\{\nu : |b(\nu)| < \infty\} = \mathbb{R}\}$, (3) for all $k \in \{1, \ldots, K\}$,

$$\{k\} = \cap_{k \in S, p_Q(S)} S,$$

where $\cap_{k \in S, p_Q(S)} S\} \triangleq \emptyset$ if $p_Q(S) = 0$ for all S that contains k. Bing et al. (2020) provides an alternative set of conditions for consistent estimation in sparse structured factor models under an anchor assumption (Arora et al., 2013).

In what follows, we will assume that if the unobserved confounder U_i is an emancipator that is determinable in theory, and if fitting the factor model to $\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i$ will also emit an estimate Z_i for U_i that is also an emancipator that is determinable in theory, then, under suitable conditions, Z_i can serve as a valid estimate for U_i in confounder adjustment.

We define the condition that U_i must satisfy in the next section—specifically, if the unobserved confounder U_i is a *multi-period confounder*.

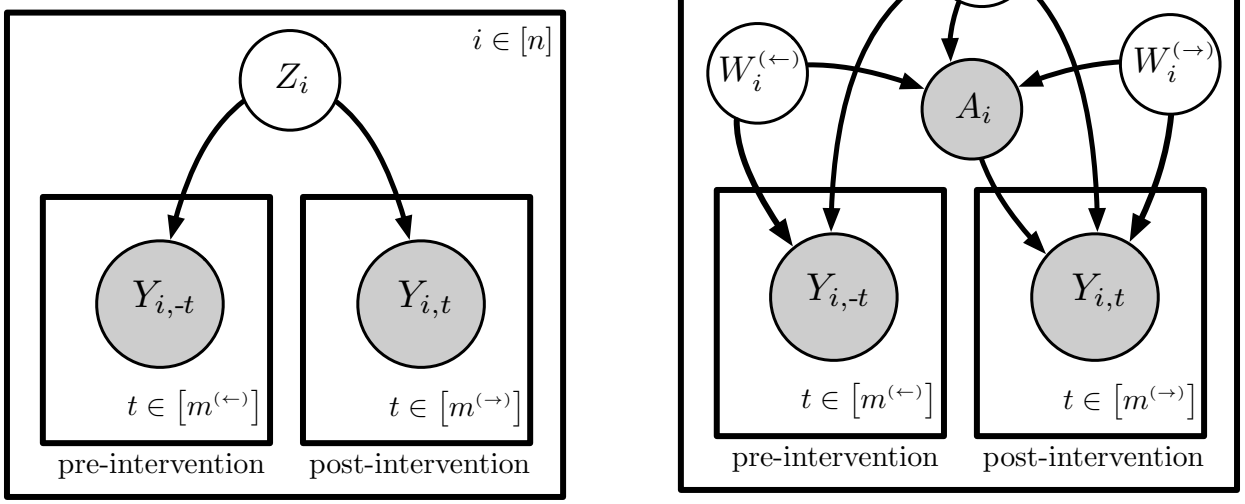
3.2 Multi-period confounders

A confounder is any variable that causally affects both the treatment and some outcomes. In panel data, when there are multiple outcomes per unit, some confounders may affect outcomes at all time steps, while others may affect only some outcomes at certain time steps. In the synthetic control setting, time steps naturally split into two fundamentally different periods: pre- and post-intervention. By extension, it is natural to distinguish two kinds of confounders: those that affect outcomes in multiple (both) periods, and those that affect outcomes in only a single period.

Definition 2 (Single- and multi-period confounders)**.** *A variable $\overleftrightarrow{W}_i$ is a* multi-period confounder *if it causally affects the treatment A_i and outcomes $Y_{i,t}$ in both the pre- and post-intervention periods. A* single-period confounder *is then a variable that causally affects A_i and outcomes in* only *the pre- or post-intervention period, denoted by $\overleftarrow{W}_i$ or $\overrightarrow{W}_i$, respectively. See also Figure 2b.*

Informally, a multi-period confounder $\overleftrightarrow{W}_i$ is a confounder that "shows its face" in both periods—it affects, and is thus reflected by, the outcomes both before and after the intervention.

Lemma 3. *If any random variable U_i is a determinable-in-theory emancipator then it captures all multi-period confounders.*



(a) Factor models for multiple control outcomes. The defining feature of factor models is conditional independence: Z_i renders all outcomes $\{Y_{i,t}\}_{t<0} \cup \{Y_{i,t}\}_{t>0}$ conditionally independent.

(b) Emancipators must pick up all multi-period confounders $\overleftrightarrow{W}_i$ but no single-period confounders $W_i^{(\leftarrow)}$ or $W_i^{(\rightarrow)}$.

Figure 2: Factor models and emancipators.

Proof. *Suppose U_i "misses"[3] some multi-period confounder $\overleftrightarrow{W}_i$. Then $\overleftrightarrow{W}_i$ is effectively unobserved and induces dependence between two or more outcomes. Conditioning on U_i therefore does not render all outcomes conditionally independent, which contradicts the definition of U_i.*

The consequence of Lemma 3 is that, if we assume U_i to be an emancipator that is determinable in theory, then it must be a multi-period confounder. We may then use factor models to estimate confounders and adjustment for them to identify causal effects from panel data. We substantiate and defend this in more detail in the next subsection.

3.3 Causal identification with a determinable-in-theory emancipator

In this section, we detail a strategy for causal identification and estimation in panel data in the presence of unobserved confounding. This strategy adjusts for unobserved confounding using an estimate Z_i of the latent variable from the data for the unobserved confounder U_i. The strategy hinges on several assumptions: (1) the stable unit treatment value assumption (SUTVA); (2) fitting the factor model produces a *emancipator* Z_i that is determinable in theory; (3) the unobserved confounder is a *multi-period confounder*; (4) overlap, also known as positivity.

This approach inherits from the classical strategy of measuring sufficient confounders, but with the difference that some confounders—the multi-period confounders—may not be directly observed. Rather, those confounders are reflected in the pre-intervention outcomes to such a degree that a

3. "miss" means some multi-period confounders are not measurable with respect to the emancipator.

variable $U_i = f(\overleftarrow{\mathbf{Y}}_i)$ can capture all of the them under suitable assumption.Below we discuss each assumption in detail, and then state the theorem that establishes identification.

Assumption 1. *(Stable unit treatment value assumption (SUTVA)) There is no interference between the units; there is only one version of the treatment for each value $a \in \mathcal{A}$.*

SUTVA is the classical assumption we make to perform causal inference on independent units (Rubin, 1980, 1990). It ensures that the potential outcome for each unit $Y_{i,t}^{(a)}$ is well-defined.

Assumption 2. *(Determinable-in-theory emancipator) The population distribution of unit i's outcomes is well-described by a parametric model θ as the marginal over some latent variable U_i which satisfies the properties of a* determinable-in-theory emancipator*, as given in Definition 1.*

Assumption 2 enforces that it is possible to fit the factor model to obtain a determinable-in-theory emancipator U_i, i.e. (1) "emancipates" all of unit i's outcomes and (2) can be "determined in theory" by i's pre-intervention outcomes. As given by Lemma 3, the first property ("emancipates") ensures that U_i "captures" all multi-period confounders. The second property ("determinable in theory") then means we can effectively observe U_i, given enough data. The second property is satisfied when (1) the distribution of all outcomes can be described by a factor model and (2) the number of pre-intervention outcomes increases to infinity, i.e., $\overleftarrow{m} \to \infty$. For example, if the distribution of all outcomes can be described by structured latent factor models, like probabilistic principal component analysis (Tipping & Bishop, 1999) or Poisson factorization (Gopalan et al., 2013, 2014, 2015), then the latent variable U_i can be determined in theory when $(n + \overleftarrow{m}) \cdot \log(n\overleftarrow{m})/(n\overleftarrow{m}) \to 0$ (Chen et al., 2017).

At a high level, Assumption 2 replaces and generalizes the main assumption of traditional synthetic controls, which is that the causal structural model can be described by a linear factor model. Instead, we assume that the causal structural model can be described by any factor model with the above properties.

Assumption 3. *(Weak unconfoundedness given unobserved but determinable-in-theory multi-period confounders) Under Assumption 2, the determinable-in-theory emancipator U_i with the observed covariates X_i satisfy weak unconfoundedness,*

$$\overleftarrow{\mathbf{Y}_i^{(a)}}, \overrightarrow{\mathbf{Y}_i^{(a)}} \perp A_i \,|\, U_i, X_i, \quad \forall a \in \mathcal{A}. \tag{15}$$

Assumption 3 requires "no unobserved single-period confounders", namely we observe any confounders that affect outcomes in only one of the two periods. This is because U_i—a determinable-in-theory emancipator—can only capture multi-period confounders; it cannot capture single-period confounders. Note that this includes "single-outcome confounders" which causally affect only a single outcome, be it control outcome or study outcome. This assumption is expressed mathematically by assuming the determinable-in-theory emancipator U_i, along with the observed covariates X_i, satisfies weak unconfoundedness (Equation (15)) (Imbens, 2000; Hirano & Imbens, 2004). We note that this assumption differs from a common identification assumption used in synthetic control (O'Neill et al., 2016; Angrist & Pischke, 2008; Kinn, 2018).

$$\overrightarrow{\mathbf{Y}_i^{(a)}} \perp A_i \,|\, X_i, \overleftarrow{\mathbf{Y}_i}, \quad \forall a \in \mathcal{A}. \tag{16}$$

Assumption 3 posits that the determinable-in-theory emancipator U_i and the covariates X_i satisfy weak unconfoundedness, while Equation (16) assumes that all pre-intervention outcomes $\overleftarrow{\mathbf{Y}_i}$ and the covariates X_i satisfy weak unconfoundedness.

Assumption 4 (Estimated determinable-in-theory emancipator)**.** *For all i, the parametric model θ in Assumption 2 admits an estimate Z_i of the latent variable U_i such that $Z_i = f(\overleftarrow{\mathbf{Y}}_i)$ is also a determinable-in-theory emancipator and consistently estimates U_i up to bijective transformation.*

Assumption 4 guarantees that the estimate Z_i is consistent for the unobserved multi-period confounder U_i up to bijective transformation; in other words, Z_i captures all information in U_i. The consistency is only up to bijective transformations because the latent variables in model θ may not be identifiable. This consistent estimate Z_i can thus be used for confounder adjustment in the place of U_i.

Assumption 5. *(Overlap, also known as positivity) For all sets $\mathcal{A}_S$ with positive measure, $P(A_i \in \mathcal{A}_S \mid X_i = x, U_i = u) > 0$ for all values of (x, u), where $U_i \overset{a.s.}{=} f(\overleftarrow{\mathbf{Y}}_i)$ as is required by Assumption 2.*

Overlap is a classical assumption required in causal inference, which often requires that all values of the pair, treatment and confounder, occur with positive probability. The overlap assumption here is similar but concerns both the unobserved multi-period confounder U_i and the covariates X_i. We emphasize that while U_i is a deterministic function of the outcomes, this does not prevent the overlap assumption from being satisfied. This is in contrast to settings with multiple treatments (Wang & Blei, 2019a) where overlap can be more difficult to satisfy; here we estimate the latent confounder from the pre-treatment outcomes, which does not interfere with the overlap condition required for the treatment.

Under these assumptions, adjusting for the estimated confounding structure Z_i can lead to unbiased estimates of average potential outcomes, including counterfactual outcomes, and thus of the ATT.

Theorem 4. *(Causal identification with the estimated confounding structure) Under Assumptions 1 to 5, controlling for the estimated confounding structure Z_i yields unbiased causal estimates,*[4]

$$\mathbb{E}\left[Y_{i,t}^{(a)}\right] = \mathbb{E}_{Z_i, X_i}\left[\mathbb{E}\left[Y_{i,t} \mid A_i = a, Z_i, X_i\right]\right] \tag{17}$$

$$\mathbb{E}\left[Y_{i,t}^{(a)} \mid A_i = a'\right] = \mathbb{E}_{Z_i, X_i \mid A_i = a'}\left[\mathbb{E}\left[Y_{i,t} \mid A_i = a, Z_i, X_i\right]\right]. \tag{18}$$

A sketch of the proof is as follows. The estimated confounding structure Z_i captures all information in the unobserved multi-period confounder U_i. Further, U_i, together with the covariate X_i, satisfy weak unconfoundedness. Hence adjusting for them with g-formula leads to unbiased estimates of the potential outcomes. The full proof of Theorem 4 is in Section B.

At a higher level, Assumptions 1 to 5 yield a setting where the class of multi-period confounders are factually unobserved, but effectively observed—they can be determined in theory by observing infinitely many pre-intervention outcomes.

3.4 Estimating causal effects with determinable-in-theory emancipators

Theorem 4 immediately leads to an algorithm that estimates causal effects by inferring determinable-in-theory emancipators from multiple control outcomes. We first construct the estimate of the unobserved multi-period confounders that all outcomes conditionally independent. Then we use

4. We note that U refers to the true latent confounder, and Z is the estimate of it. We avoid writing $\hat{U}$ because U is generally only identifiable up to an unknown bijective transformation. So we in generally cannot recover U directly, only an equivalent representation Z that preserves its effect on causal adjustment. Using a distinct symbol (Z) emphasizes this identifiability limitation and avoids suggesting that we are estimating U in a pointwise or coordinate-wise sense.

that estimate in a downstream causal in **??**, show how the synthetic control method is a special case.

Estimating the determinable-in-theory emancipator Z_i. If Assumption 3 holds, then it is sufficient to perform confounder adjustment with any determinable-in-theory emancipator Z_i (and observed covariates); the first key characteristic of such a variable is that it "emancipates" all of unit i's outcomes,

$$P\left(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i \mid Z_i\right) = \left[\prod_{t=1}^{\overleftarrow{m}} P\left(Y_{i,-t} \mid Z_i\right)\right] \left[\prod_{t=1}^{\overrightarrow{m}} P\left(Y_{i,t} \mid Z_i\right)\right]. \tag{19}$$

We enforce this conditional independence by finding a factor model θ that captures the distribution of the outcomes $P(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i) = \int P_\theta(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i, Z_i)\, dZ_i$. A factor model posits a generative process such that

$$Z_i \overset{\text{ind.}}{\sim} P(Z_i), \qquad \forall i, \tag{20}$$

$$Y_{i,t} \mid Z_i \overset{\text{ind.}}{\sim} P(Y_i \mid Z_i, \theta_t) \qquad \forall i, t, \tag{21}$$

where $\boldsymbol{\theta} = (\theta_t)_t$ are parameters of the factor model. In particular, the latent variable Z_i renders the outcomes conditionally independent.[5]

We fit the factor model by optimizing the likelihood of the factor model over the parameters

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \int P(\overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i, Z_i \mid \boldsymbol{\theta})\, dZ_i, \tag{22}$$

and then obtain a point estimate of the latent variable Z_i:

$$\hat{z}_i = \mathbb{E}\left[Z_i \mid \overleftarrow{\mathbf{Y}}_i, \hat{\boldsymbol{\theta}}\right]. \tag{23}$$

When $\overleftarrow{m} \to \infty$—i.e., the number of pre-intervention outcomes grows large—the posterior distribution converges to a point mass $P(Z_i \mid \overleftarrow{\mathbf{Y}}_i, \boldsymbol{\theta}) \overset{\text{d}}{\to} \delta\left(f_{\boldsymbol{\theta}}\left(\overleftarrow{\mathbf{Y}}_i\right)\right)$ for some function $f_{\boldsymbol{\theta}}$. The point estimate will thus also converge—i.e. $\hat{z}_i \overset{\text{a.s.}}{\to} f_{\boldsymbol{\theta}}\left(\overleftarrow{\mathbf{Y}}_i\right)$.

Many classical probabilistic models fall into the class of factor models, including probabilistic PCA (Tipping & Bishop, 1999), Poisson factorization (Gopalan et al., 2013, 2014, 2015), mixture models (McLachlan & Basford, 1988), topic models (Blei et al., 2003), and deep exponential families (Ranganath et al., 2015). As an example, probabilistic PCA posits that

$$Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_Z^2 \cdot \mathrm{I}_K), \qquad \forall i, \tag{24}$$

$$Y_{i,t} \mid Z_i \overset{\text{ind.}}{\sim} \mathcal{N}(Z_i^\top \theta_t, \sigma_Y^2), \qquad \forall i, t. \tag{25}$$

5. Theoretically, this factor model only needs to hold true for pre-intervention outcomes for all units. In practice, when a large number of pre-treatment outcomes are available, excluding the treated units from this stage typically makes little difference. However, in finite samples, especially when only a moderate number of pre-treatment outcomes are available, incorporating post-treatment outcomes during latent factor modeling can help focus the model on the components most relevant for causal adjustment. This is because, for a variable to act as a confounder, it must affect both treatment assignment and outcomes. Prioritizing such factors in the representation often improves the efficiency of causal adjustment in practice.

A reader might ask: Why fitting the factor model to the units-by-time matrix as opposed to a time-by-units matrix? The choice reflects a core idea behind synthetic control: we are typically interested in estimating treatment effects averaged over units, not over time. This motivates modeling units as the independent entities, each influenced by time-invariant latent confounders. The pre-treatment time periods then act as negative control outcomes; they help us learn each unit's latent characteristics, which are then used to predict their counterfactual outcomes under no treatment.

If we were to transpose the matrix and fit a latent variable model where the units are conditionally independent given latent time factors, this would implicitly frame the goal as estimating treatment effects averaged over time, assuming that some latent unit-invariant (but time-specific) factors explain all variation. That structure might make sense in settings where some time-average causal effects on individual units may be of interest. But in synthetic control settings, where the goal is typically to infer the effect of treatment averaged over a group of units post-treatment, this framing becomes less natural.

Finally, merely fitting a factor model to the outcomes does not license Z_i for confounder adjustment in the place of U_i. Fitting a factor model does not guarantee that its latent variable Z_i satisfies the conditional independence requirement (Equation (19)). The factor model may also be wrong, which can introduce bias in the downstream causal estimation.

To this end, we further check that the fitted factor model does capture the distribution of the outcomes well. The checking step cannot guarantee that the factor model is correct when it passes the check. However, it can reject wrong factor models that poorly fit the data.

Many existing algorithms can perform this check. Here we employ a predictive check (**?**). We randomly hold out some outcomes for each user and fit the factor model. Then we generate the heldout entries using the fitted factor model. If the generated heldout entries are indistinguishable (in terms of log likelihood values) from the observed values in these heldout entries, we claim the factor model passes the predictive check and can capture the distribution of the outcomes.

Note we infer the determinable-in-theory emancipator by only looking at the outcomes—it does not involve the treatment statuses. While joint inference (e.g., full Bayesian modeling of both latent variables and outcomes) is compatible with the identifiability theory, we adopt a two-stage inference approach in part to align with standard practice in synthetic control. We restrict the latent variable estimation to pre-treatment outcomes to help avoid using post-treatment information in a way that would compromise causal validity.

Confounder adjustment with the determinable-in-theory emancipator. After obtaining a determinable-in-theory emancipator $\{\hat{z}_i\}_{i=1}^n$ for each data point, we use them in a downstream causal inference, as though they were observed confounders. For example, we can compute the conditional expected potential outcome $\mathbb{E}\left[Y_{i,t}^{(a)} \mid A_i = a'\right]$ by fitting an outcome model

$$\mathbb{E}\left[Y_{i,t} \mid A_i = a, Z_i, X_i\right] = \beta_0 + \beta_A \cdot a + \beta_Z \cdot Z_i + \beta_X \cdot X_i \tag{26}$$

and applying the adjustment

$$\mathbb{E}_{Z_i, X_i \mid A_i = a'}\left[\mathbb{E}\left[Y_{i,t} \mid A_i = a, Z_i, X_i\right]\right] \approx \beta_0 + \beta_A \cdot a + \beta_Z \cdot \frac{\sum_{i:\, A_i = a'} Z_i}{\sum_{i=1}^n \mathbb{I}\{A_i = a'\}} + \beta_X \cdot \frac{\sum_{i:\, A_i = a'} X_i}{\sum_{i=1}^n \mathbb{I}\{A_i = a'\}}. \tag{27}$$

Algorithm 1: Generalization of the synthetic control method to nonlinear factor models

Input: a data set of units' treatment status, pre- and post-intervention outcomes, and covariates $\left(a_i, \overleftarrow{\mathbf{y}_i}, \overrightarrow{\mathbf{y}_i}, x_i\right)_{i=1}^n$

Output: the conditional average potential outcome $\mathbb{E}\left[Y_{i,t}^{(a)} \mid A_i = a'\right] \quad \forall a \in \mathcal{A}$

repeat

  choose a factor model from the class described in Section 3.4

  fit the model to the outcomes $\left(\overleftarrow{\mathbf{y}_i}, \overrightarrow{\mathbf{y}_i}\right)_{i=1}^n$

  check the fitted model $\hat{M}$

until *the factor model check is satisfactory*

foreach *unit i* **do**

  calculate $\hat{z}_i = \mathbb{E}_{\hat{M}}\left[Z_i \mid \overleftarrow{\mathbf{Y}_i}\right]$.

end

repeat

  choose a study outcome model, e.g. Equation (26)

  fit the outcome model to the augmented dataset $\{(a_i, y_i, \hat{z}_i, x_i)\},\ i = 1, \ldots, n$

  check the fitted outcome model

until *the outcome check is satisfactory*

estimate the conditional average potential outcome $\mathbb{E}\left[Y_{i,t}^{(a)} \mid A_i = a'\right]$ by Equation (27), Equation (28), or other adjustment methods

Alternatively, we may use the inverse propensity score weighting to calculate $\mathbb{E}\left[Y_{i,t}^{(a)} \mid A_i = a'\right]$

$$\sum_{i:A_i=a} Y_{i,t} \cdot \frac{1}{\sum_{i=1}^n \mathbb{I}\{A_i = a'\}} \cdot \frac{P(A_i = a' \mid Z_i, X_i)}{P(A_i = a \mid Z_i, X_i)}. \tag{28}$$

Both the adjustment formula and inverse propensity score weighting are standard confounder adjustment methods. We may use other adjustment methods, e.g. doubly-robust estimators or double machine learning; we can estimate other quantities too, e.g. conditional ATE and individual treatment effects. The only twist is that we join the covariates X_i with the determinable-in-theory emancipator Z_i, treating both as confounders we adjust for.

Algorithm 1 presents the full algorithm. In summary, this strategy helps adjust for unobserved multi-period confounders U_i, those shared between both pre- and post-intervention outcomes. Thanks to the identification theorem of Theorem 4, and subject to the assumptions, this algorithm provides unbiased causal inferences.

Uncertainty quantification of Algorithm 1. To quantify the sampling uncertainty of Algorithm 1 (Equations (27) and (28)), we employ the bootstrap method (Efron & Tibshirani, 1994). We resample the per-unit data points $(A_i, \overleftarrow{\mathbf{Y}_i}, \overrightarrow{\mathbf{Y}_i})$ with replacement. For each resampled dataset, we calculate the causal estimate from Algorithm 1 a confidence interval. We evaluate the coverage of bootstrap uncertainty estimates in Section 5.

3.5 A graphical interpretation of the synthetic control assumptions

In this section, we provide a graphical interpretation of the identification assumptions, including both the linear structural model described in Equations (8) to (10) presented in Section 3.4 and the alternative assumptions proposed in Section 3.3 for nonlinear factor models (Equation (12)). We show that the conditional independence assumption of factor models and the weak unconfoundedness assumption can be posed purely in terms of the graphical structure of the assumed structural model.

Pre-emptive disclaimer. The graphical models in this section are *probabilistic* graphical models (e.g., Koller & Friedman (2009)), not *causal* graphical models (Pearl, 2009)—they encode the conditional dependencies between variables in a joint distribution, and nothing more.

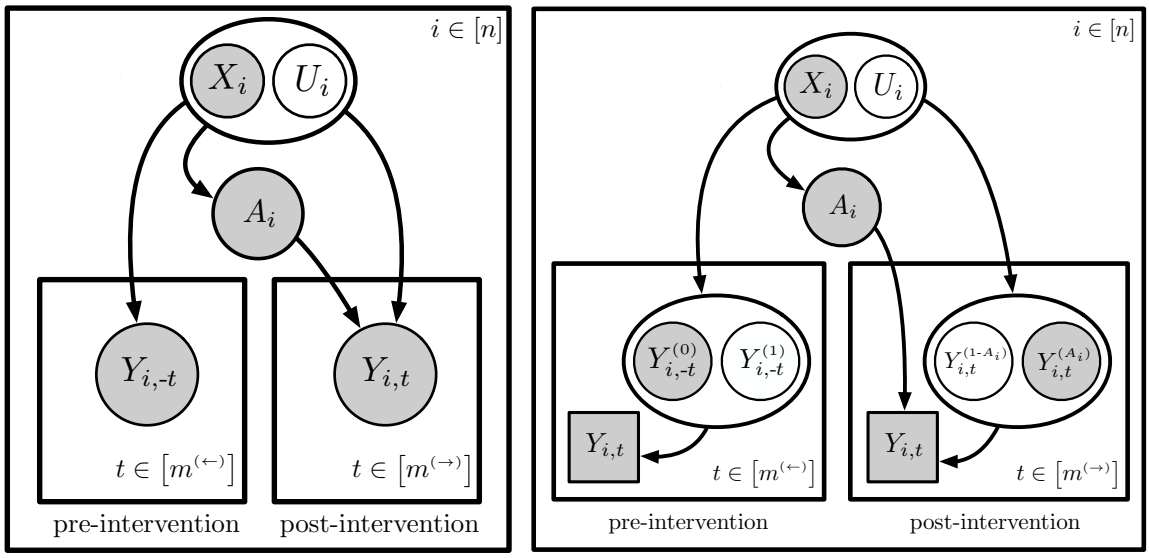
The basic graphical model. Figure 3a describes the basic dependence relations between the observed outcomes $Y_{i,t}$, the treatment status A_i, the observed covariates X_i, and unobserved attributes U_i. The plate notation denotes repeated variables. The outer plate denotes repetition across units $i \in [n] \equiv \{1, \ldots, n\}$, while the left and right inner plates denote repetition across pre-intervention ($t < 0$) and post-intervention ($t > 0$) time steps, respectively. Shaded variables are observed while unshaded nodes are latent. The super node notation for X_i and U_i denotes that these variables share the same set of edges to other nodes in the graph. We adopt this notation to avoid cluttering the figure with repeated edges, but also to emphasize that X_i and U_i are conceptually similar—both variables represent attributes of the unit i that are potentially prognostic of both i's treatment status A_i and observed outcomes $Y_{i,t}$.

The main purpose of this graphical model is to highlight that A_i only affects (i.e., has arrows to) the post-intervention outcomes while U_i affects both the pre- and post-intervention outcomes.

The augmented graphical model. Figure 3b depicts the same model, but augmented to include the potential outcomes $Y_{i,t}^{(0)}, Y_{i,t}^{(1)}$. We use square nodes to indicate deterministic variables. The outcomes $Y_{i,t}$ are deterministic functions of potential outcomes and A_i—i.e., $Y_{i,t} = Y_{i,t}^{(A_i \cdot \mathbb{I}(t>0))}$—and thus appear as square nodes in this graph. As in the other graph, A_i only affects the post-intervention outcomes. Specifically, A_i selects which of the potential outcomes is observed; this is denoted by $Y_{i,t}^{(A_i)}$ being shaded, since it equals $Y_{i,t}$ and is thus observed. In the pre-intervention period, $Y_{i,t}$ always coincides with $Y_{i,t}^{(0)}$; as a result, $Y_{i,t}^{(0)}$ is shaded, and there is no arrow from A_i to $Y_{i,t}$.

The main purpose of this graphical model is to emphasize that $Y_{i,t}^{(0)}$ is observed for all units in the pre-intervention period, while only observed for the donor units in the post-intervention period. Moreover, X_i, U_i renders A_i independent of $Y_{i,t}^{(1-A_i)}, Y_{i,t}^{(A_i)}$ in the augmented graphical model; this conditional independence coincide with weak unconfoundedness in Assumption 3.

The synthetic control method as a special case of Algorithm 1. We have seen that the linear structural model satisfies the identification condition above. Here we will see that the synthetic control method is a special case of Algorithm 1. Specifically, the synthetic control method obtains the unit weights by solving a system of (population) balancing equations; the inverse probability weights induced by Algorithm 1 will satisfy the same population balancing equations. Taking this perspective, the synthetic control method can be viewed as implicitly calculating the inverse probability weights when these balancing equations have a unique solution.



(a) The basic graphical model. **(b)** The augmented graphical model.

Figure 3: Super nodes (e.g., the node for (X_i, U_i)) denote variables with the exact same set of in- and out-going edges. Multiple pre-intervention outcomes with shared unobserved confounding. (Shaded nodes are observed; unshaded ones are latent.)

Suppose the dataset $\{A_i, \overleftarrow{\mathbf{Y}}_i, \overrightarrow{\mathbf{Y}}_i\}_{i=1}^n$ is generated by the linear structural model (Equation (8)) with the unobserved confounders U_i and observed covariates X_i. This connection relies on a few observations:

1. Applying Algorithm 1 to the dataset, we obtain an estimated confounding structure Z_i that is consistent for U_i. Together with covariates X_i, the estimated confounding structure Z_i satisfies Assumptions 2 to 4.

2. To estimate $\mathbb{E}\left[Y_{i,t}^{(0)} \mid A_i = 1\right]$, we calculate the inverse probability weighting estimator (Horvitz & Thompson, 1952)

$$\sum_{A_i=0} \overrightarrow{\mathbf{Y}}_i \cdot w_i^{\text{IPW}}, \tag{29}$$

where the weights are

$$w_i^{\text{IPW}} = \frac{1}{nP(A_i = 1)} \cdot \frac{P(A_i = 1 \mid Z_i, X_i)}{P(A_i = 0 \mid Z_i, X_i)}. \tag{30}$$

These weights can balance the pre-treatment outcomes because of Assumptions 3 and 4, i.e. these outcomes are weakly unconfounded given U_i, X_i (Rosenbaum & Rubin, 1983), and Z_i consistently estimates U_i, i.e.

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} Y_{i,t}\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^n \mathbb{I}\{A_i = 1\}} \sum_{A_i=1} Y_{i,t}\right], j = 1, \ldots, m. \tag{31}$$

Roughly, these observations imply that Algorithm 1 finds the unobserved confounder U_i by fitting probabilistic factor models. The resulting inverse probability weights hence satisfy the synthetic

control constraints because U_i, X_i render the outcomes weakly unconfounded. We formally state these connections in **??** and **??**.

This connection between synthetic control and Algorithm 1 also generalizes to other variants of synthetic control methods. For example, the same connection also holds with robust synthetic controls and its multi-dimensional variant (Amjad et al., 2018, 2019). The exact same argument applies if we consider all outcomes $Y_i, \widetilde{Y}_i^1, \ldots, \widetilde{Y}_i^m$ being L–dimensional vectors ($L > 1$).

Proposition A.2 in Ben-Michael et al. (2021) provides a similar result that connects the synthetic-control estimator to inverse-propensity-score weighting. The main difference is that their result relies on inverse-propensity-score weights conditioned on the full vector of pre-treatment outcomes, $\overleftarrow{\mathbf{Y}}_i$, whereas **??** uses inverse-propensity-score weights conditioned solely on the determinable-in-theory emancipator Z_i (Algorithm 1).

We remark that Algorithm 1 relates closely to Xu (2017) and Gobillon & Magnac (2016) as two stage algorithms for panel data. The difference lies in that Algorithm 1 explicitly uncovers potential unobserved confounders by fitting probabilistic factor models, while the others optimize weights, which implicitly handles unobserved confounders. The setting of many control outcomes ($m \to \infty$) also closely relates to Ferman (2019), who studies the limiting properties of the synthetic control method but with infinitely many pre-treatment periods. However, their results focus on linear structural models while we study more general settings.

Finally, we note that in our reinterpretation of synthetic control, the treatment A_i appears as a node in a graphical model and thus might appear to be a random variable, which would depart from canonical work on synthetic control that views the treatment as deterministic. Our framework assumes A_i is always observed, as denoted by the variable's shading in the graphical model, and is consistent with either a deterministic or random interpretation of A_i, either of which may be appropriate depending on the context. For example, in the NFL example, it may be appropriate to view the treatment status as random, as COVID-19 policies were highly variable across even neighboring counties in the US (Hamad et al., 2022; Chiang et al., 2025). However, the proposed framework makes no stipulation that the treatment assignment mechanism be either random or deterministic.

4. Replication and re-analysis of García Bulle et al. (2022): Did re-opening NFL stadiums impact the spread of COVID-19?

We return now to the running example in Section 1 on whether the National Football League (NFL)'s decision to re-open stadiums to fans in September 2020 had an impact on the spread of COVID-19. This example comes from the study of García Bulle et al. (2022) who collected panel data on NFL attendance and COVID-19 case counts by US county and applied the method of robust synthetic control (Amjad et al., 2018). Their results suggest that the re-introduction of fans to NFL stadiums had a negligible subsequent impact on rates of COVID-19 in the US counties where fans mainly lived.

In this section, we first provide a replication of their study, and then re-analyze their data using the proposed Algorithm 1. We find that the results obtained using the proposed method qualitatively agree with the conclusion that re-opening stadiums had negligible impact. The proposed method generally estimates counterfactual rates of COVID-19 that are even closer to the factual rates than those estimated by rSC for both stadiums that were re-opened and "placebo" stadiums that were not.

Moreover, we find that the proposed method is robust to violations in the data of rSC's assumptions, which García Bulle et al. (2022) highlight as leading to strange counterfactual predictions for particular counties. We replicate those strange predictions and find that, unlike rSC, the proposed Algorithm 1 generates sensible counterfactual predictions in these cases. Our results suggest that allowing for a broader class of factor models can have a meaningful impact in cases where the data exhibit violations of the assumptions of simple linear factor models.

4.1 Data description

The data of García Bulle et al. (2022) combine daily COVID case counts by US county with attendance records from the NFL that say from which US counties fans originated. They compiled separate data sets for 31 NFL teams[6] and analyzed each one independently.

For each NFL team, the counties supplying 10% or more of the fan base are designated the targets. The donors are then all counties that

1. supply less than 1% of the fan base,
2. are in the same state as the respective stadium, and
3. did not undergo any idiosyncratic shocks during the study period (April–December 2020).

Counties that supply between 1–10% fan base are neither targets nor donors; they are designated as "buffer" counties and excluded from the analysis.

4.2 Replication of García Bulle et al. (2022)

Before applying our proposed method, we first attempted to replicate the counterfactual predictions that García Bulle et al. (2022) obtained for the target counties in each of the 31 data sets. They obtained counterfactual predictions using the method of robust synthetic control (rSC) which projects the donors' outcomes into a low-dimensional subspace using principal components analysis (PCA). An important setting in applying rSC is the dimensionality of PCA. We follow the procedure they outline (see their *Materials and Methods*) to select this dimensionality and obtain the same value they do in all but two cases[7]. Using these values, and we then applied our implementation of rSC and were able to reproduce counterfactual predictions that closely matched theirs by visual inspection (compare their Figure 2 to our Figure 5).

4.3 Re-analysis with the proposed Algorithm 1

Having replicated their setup, we then applied our proposed methodology to each of the 31 data sets. We provide a detailed illustration in this subsection of the steps outlined in Section 3. We then report the counterfactual predictions that the proposed methodology obtains and compare them to those obtained by García Bulle et al. (2022).

4.3.1 Choosing a factor model with heldout predictive checks

Following Algorithm 1, the first step in the proposed methodology is to find a probabilistic factor model that fits the pre-intervention data matrix.

6. This is all NFL teams except the Arizona Cardinals for whom attendance data was unavailable.
7. We found a small discrepancy between the procedure they outline in the paper, and their implementation, which we confirmed with the authors and correct in our own replication.

PPCA. We first considered probabilistic principal components analysis (Tipping & Bishop, 1999), a linear factor model that assumes the following generative process

$$\theta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathrm{I}_K), \qquad \forall t, \tag{32}$$

$$Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathrm{I}_K), \qquad \forall i, \tag{33}$$

$$Y_{i,t} \,|\, Z_i,\, \theta_t \overset{\text{ind.}}{\sim} \mathcal{N}(Z_i^\top \theta_t,\, \sigma_Y^2) \qquad \forall i, t. \tag{34}$$

In what follows, we adopt the following notation: $\boldsymbol{Z} \equiv (Z_1, \dots Z_n)$ is the $n \times K$ matrix of latent variables, $\overleftarrow{\mathbf{Y}} \equiv (\overleftarrow{\mathbf{Y}}_1, \dots, \overleftarrow{\mathbf{Y}}_n)$ is the $n \times \overleftarrow{m}$ matrix of pre-intervention outcomes, and $\overleftarrow{\Theta} \equiv (\theta_t)_{t<0}$ is the $\overleftarrow{m} \times K$ matrix of pre-intervention latent variables.

NUTS. We implemented this model in Pyro (Bingham et al., 2019), a probabilistic programming framework in Python, and fit it using Pyro's implementation of the No U-Turn Sampler (NUTS) (Hoffman et al., 2014). NUTS is a form of Markov Chain Monte Carlo (MCMC) which returns samples of the latent variables from their posterior distribution. Using this notation, the target of inference is the posterior distribution $P(\boldsymbol{Z},\, \overleftarrow{\Theta} \mid \overleftarrow{\mathbf{Y}})$, which MCMC approximates with a set of R samples: $\{\boldsymbol{Z}_r,\, \overleftarrow{\Theta}_r\}_{r=1}^R$, each drawn from the posterior.

Heldout-PCs. We assessed model fit using a heldout predictive check (Heldout-PC) (**?**). This involves splitting $\overleftarrow{\mathbf{Y}}$ into $\overleftarrow{\mathbf{Y}}^{(\text{obs})}$ and $\overleftarrow{\mathbf{Y}}^{(\text{new})}$, where $\overleftarrow{\mathbf{Y}}^{(\text{obs})}$ is a subset of entries in the pre-intervention matrix and $\overleftarrow{\mathbf{Y}}^{(\text{new})}$ is its complement. We do this by setting $\overleftarrow{\mathbf{Y}}^{(\text{new})}$ to be a randomly-selected 10% of entries throughout the matrix whose values are masked during training. For each of the 31 data sets (each corresponding to an NFL team), we randomly generate 5 different masks and repeat our experiments for each.

For a given data set and mask, we fit the factor model to $\overleftarrow{\mathbf{Y}}^{(\text{obs})}$ by running MCMC and collecting a set of R posterior samples, each drawn from $(\boldsymbol{Z}_r,\, \overleftarrow{\Theta}_r) \sim P(\boldsymbol{Z},\, \overleftarrow{\Theta} \mid \overleftarrow{\mathbf{Y}}^{(\text{obs})})$. We then simulate R "replicate" data sets, each one sampled from the conditional likelihood of the heldout data, conditioned on one sample of the latent variables $\overleftarrow{\mathbf{Y}}_r^{(\text{rep})} \sim P(\overleftarrow{\mathbf{Y}}^{(\text{new})} \mid \overleftarrow{\Theta}_r, \mathbf{Z}_r)$. The set of $\{\overleftarrow{\mathbf{Y}}_r^{(\text{rep})}\}_{r=1}^R$ thus collectively constitutes a discrete approximation to the posterior predictive distribution $P(\overleftarrow{\mathbf{Y}}^{(\text{new})} \mid \overleftarrow{\mathbf{Y}}^{(\text{obs})})$.

A Heldout-PC evaluates model fit by comparing the true heldout data against the (simulated) posterior predictive distribution. **?** advocate for computing the following empirical p-value:

$$p_{\text{pop}} = \tfrac{1}{R} \sum_{r=1}^{R} \mathbb{I}\left[d\big(\overleftarrow{\mathbf{Y}}_r^{(\text{rep})}, \boldsymbol{Z}_r,\, \overleftarrow{\Theta}_r\big) \geq d\big(\overleftarrow{\mathbf{Y}}^{(\text{new})}, \boldsymbol{Z}_r,\, \overleftarrow{\Theta}_r\big)\right] \tag{35}$$

where the realized diagnostic $d(\cdot)$ is the negative log-likelihood

$$d\big(\overleftarrow{\mathbf{Y}}^{(\text{new})}, \boldsymbol{Z}_r,\, \overleftarrow{\Theta}_r\big) = -\log P(\overleftarrow{\mathbf{Y}}^{(\text{new})} \mid \boldsymbol{Z}_r,\, \overleftarrow{\Theta}_r) \tag{36}$$

If the model is well-specified p_{pop} should be uniformly distributed. Therefore, a policy that rejects a model unless it has $p_{\text{pop}} \in [\frac{\alpha}{2},\, 1 - \frac{\alpha}{2}]$ will have a false rejection rate of $\alpha \in (0, 1)$. Throughout we adopt a standard threshold of $\alpha = 0.05$.

PPCA results. For each county and mask, we fit PPCA using a range of values for $K \in \{5, 10, 15, 25, K^\star\}$, where $K^\star$ is the value selected for PCA in our replication of García Bulle et al. (2022).

Our results clearly indicate that PPCA is a poor model class for the data. For almost all combinations of counties, masks, and values of K, we find p_{pop}-values of 1.0, indicating that the simulated posterior predictive distribution is unlike the true distribution of heldout data.

Upon reflection, this is perhaps unsurprising: the discrete case counts violate PPCA's main assumption that the data are Gaussian-distributed.

GAP. We next considered a factor model tailored towards discrete data, the gamma-Poisson (GaP) factor model (Canny, 2004; Cemgil, 2009), which assumes

$$\theta_{t,k} \overset{\text{iid}}{\sim} \text{Gamma}(a_0, b_0), \qquad \forall t, k \tag{37}$$

$$Z_{i,k} \overset{\text{iid}}{\sim} \text{Gamma}(a_0, b_0), \qquad \forall i, k \tag{38}$$

$$Y_{i,t} \mid Z_i, \theta_t \overset{\text{ind.}}{\sim} \text{Pois}\left(Z_i^\top \theta_t\right) \qquad \forall i, t. \tag{39}$$

This model is similar to PPCA in every way except that it makes probabilistic assumptions which are not trivially violated by counts.

GAP results. We implemented GAP in Pyro and fit it with NUTS using the exact same procedure as outline above for PPCA. We fit it using the same set of K values to each of the the same data set and mask combinations.

Unlike PPCA, we find that for almost all of the 31 data sets, there are one or more values of K for with which the GAP model obtains values of $p_{\text{pop}} \in (\frac{\alpha}{2}, 1-\frac{\alpha}{2})$, suggesting that the simulated posterior predictive distribution is not obviously different than the true distribution of heldout data.

This is not always the case: for two data sets, corresponding to Carolina Panthers and San Fransisco 49ers, there are no values of K for which the GAP model produces non-extreme p_{pop}-values. Moreover, for the data sets and values of K for which the GAP model does obtain non-extreme p_{pop}-values, it rarely does so for all masks.

That said, the results indicate that GAP substantially improves in fit and should be preferred to generate counterfactual predictions over PPCA. One might continue to refine the model class to address potential mismatch—e.g., by considering a negative binomial likelihood, or by introducing time-varying latent factors. However, for the purpose of this case study, we continue to the next step having selected GAP as our factor model.

4.3.2 Choosing a study outcome model

Having found a satisfactory factor model, we move to the next step in Algorithm 1: choosing a study outcome model. For this, we seek a model for the conditional expectation,

$$\mathbb{E}\left[\overrightarrow{\mathbf{Y}_i} \mid A_i = a, Z_i, X_i\right] = \beta_0 + \beta_A \cdot a + \beta_Z^\top Z_i + \beta_X^\top X_i \tag{40}$$

where the dependent variable $\overrightarrow{\mathbf{Y}_i}$ is the $\overrightarrow{m}$-dimensional vector of post-intervention outcomes. For simplicity, we use Ridge regression. We also experimented with other linear regressors, as well as a handful of non-linear regressors, but found little variability in the results.

Fitting the study outcome model amounts to regressing $\overrightarrow{\mathbf{Y}}_i$ on A_i, $\widehat{Z}_i$, and X_i, using all counties i. While A_i and covariates X_i are observed, Z_i is latent, and we must use an estimate $\widehat{Z}_i$ that comes from fitting the factor model. However, fitting the factor model using MCMC does not yield a single point estimate but rather a collection of posterior samples $\left(Z_i^{(s)}\right)_{s=1}^{S}$. One might consider fitting the study outcome model using the posterior mean $\widehat{Z}_i = \frac{1}{S}\sum_{s=1}^{S} Z_i^{(s)}$ or some other aggregation of the posterior samples. However, due to the potential for "label-switching" (Stephens, 2000), aggregating across the posterior samples may not be well-defined. Moreover, doing so would fail to propagate the factor model's posterior uncertainty.

Instead, we fit a different study outcome model for every posterior sample s. For each s, define the fitted regression function for each unit i, as a function of $a \in \{0, 1\}$, to be:

$$\widehat{f}_i^{(s)}(a) \equiv \widehat{\mathbb{E}}\left[\overrightarrow{\mathbf{Y}}_i \mid A_i = a,\, Z_i = Z_i^{(s)}, X_i\right] \equiv \widehat{\beta}_0 + \widehat{\beta}_A \cdot a + \widehat{\beta}_{Z^{(s)}}^{\top} Z_i^{(s)} + \widehat{\beta}_X^{\top} X_i. \tag{41}$$

4.3.3 Counterfactual predictions

With a fitted study outcome model for each posterior sample, an estimate of average potential outcome at post-intervention time $t > 0$ is:

$$\widehat{\mathbb{E}}_s\left[Y_{i,t}^{(a)}\right] = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{i,t}^{(s)}(a) \tag{42}$$

where the expectation is with respect to the population of units i. We can regard the variability across s as preserving the Bayesian posterior uncertainty from the factor model. At a given post-intervention time $t > 0$, our estimate of the total treatment effect on the treated (TTT) is then

$$\widehat{\mathrm{TTT}}_s(t) = \sum_{i:A_i=1}\left[Y_{i,t} - \widehat{f}_{i,t}^{(s)}(0)\right] \tag{43}$$

which estimates the total number of COVID-19 cases in county i, above (or below) the factual amount, that would (or would not) have occurred at t time steps after the intervention. We can similarly estimate the total treatment effect on the untreated (TTU) by:

$$\widehat{\mathrm{TTU}}_s(t) = \sum_{i:A_i=0}\left[\widehat{f}_{i,t}^{(s)}(1) - Y_{i,t}\right] \tag{44}$$

Again, we regard variability across s as preserving a measure of Bayesian uncertainty. However, we can obtain a point estimates using the posterior expectation—e.g., for the TTT:

$$\widehat{\mathrm{TTT}}(t) = \frac{1}{S}\sum_{s=1}^{S}\sum_{i:A_i=1}\left[\widehat{f}_{i,t}^{(s)}(1) - \widehat{f}_{i,t}^{(s)}(0)\right]. \tag{45}$$

4.4 Results

Figure 4 visualizes estimates of the TTT (top) and TTU (bottom) across post-intervention time step t. The box-and-whisker plots denote variability across posterior samples. The boxes depict the interquartile range and the whiskers cover the 95% credible interval. As expected, posterior uncertainty increases considerably as we move away from the intervention time $t = 0$. For all time steps, a null total treatment effect of 0 is contained well within the 95% credible interval. As one

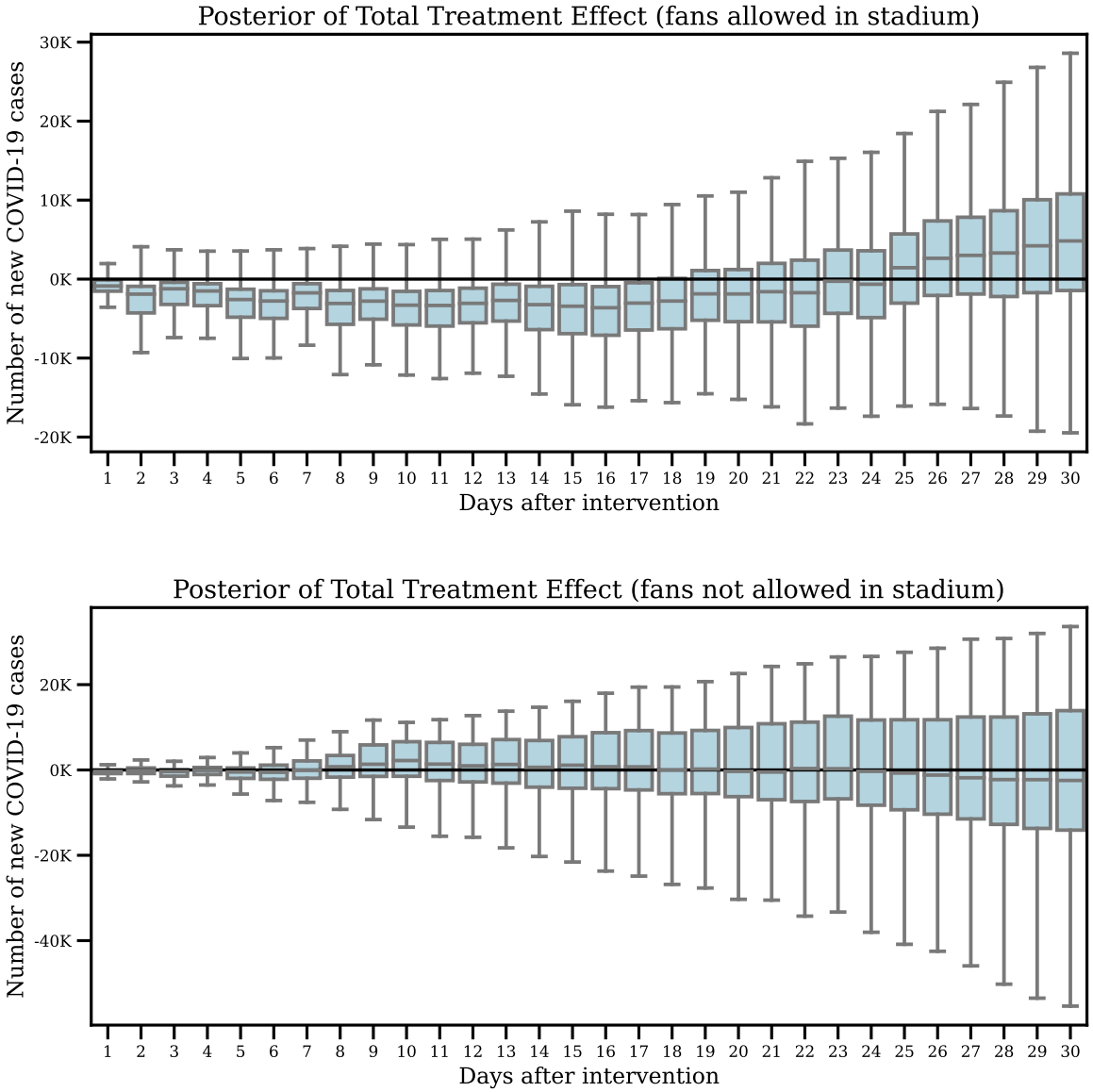


Figure 4: Re-analysis of total treatment effect by García Bulle et al. (2022).

might expect, this is also true for the TTU, for which the posterior centered on 0 for almost all time steps. These results are entirely consistent with those of García Bulle et al. (2022), who also found that allowing fans back into the stadiums did not lead to any excess COVID-19 cases.

In Figure 5, we also visualize the regression function $\widehat{f}_{i,t}^{(s)}(0)$ over time for both counties i that allowed fans in $A_i = 1$ and those that did not $A_i = 0$. For the former, the function constitutes a counterfactual trajectory of COVID-19 case counts. For the latter, it constitutes a kind of "placebo" trajectory, as we expect the function to be close to the factual outcomes. We visualize the posterior median of $\widehat{f}_{i,t}^{(s)}(0)$ as a dotted blue line, along with a band denoting the posterior interquartile range. For reference, we also visualize as dotted green lines the corresponding counterfactual and placebo trajectories obtained using our implementation of the approach outlined by García Bulle

et al. (2022); visual inspection shows little difference between these and the ones reported in their Figure 2, suggesting that our replication is correct. We also provide the factual trajectories in orange.

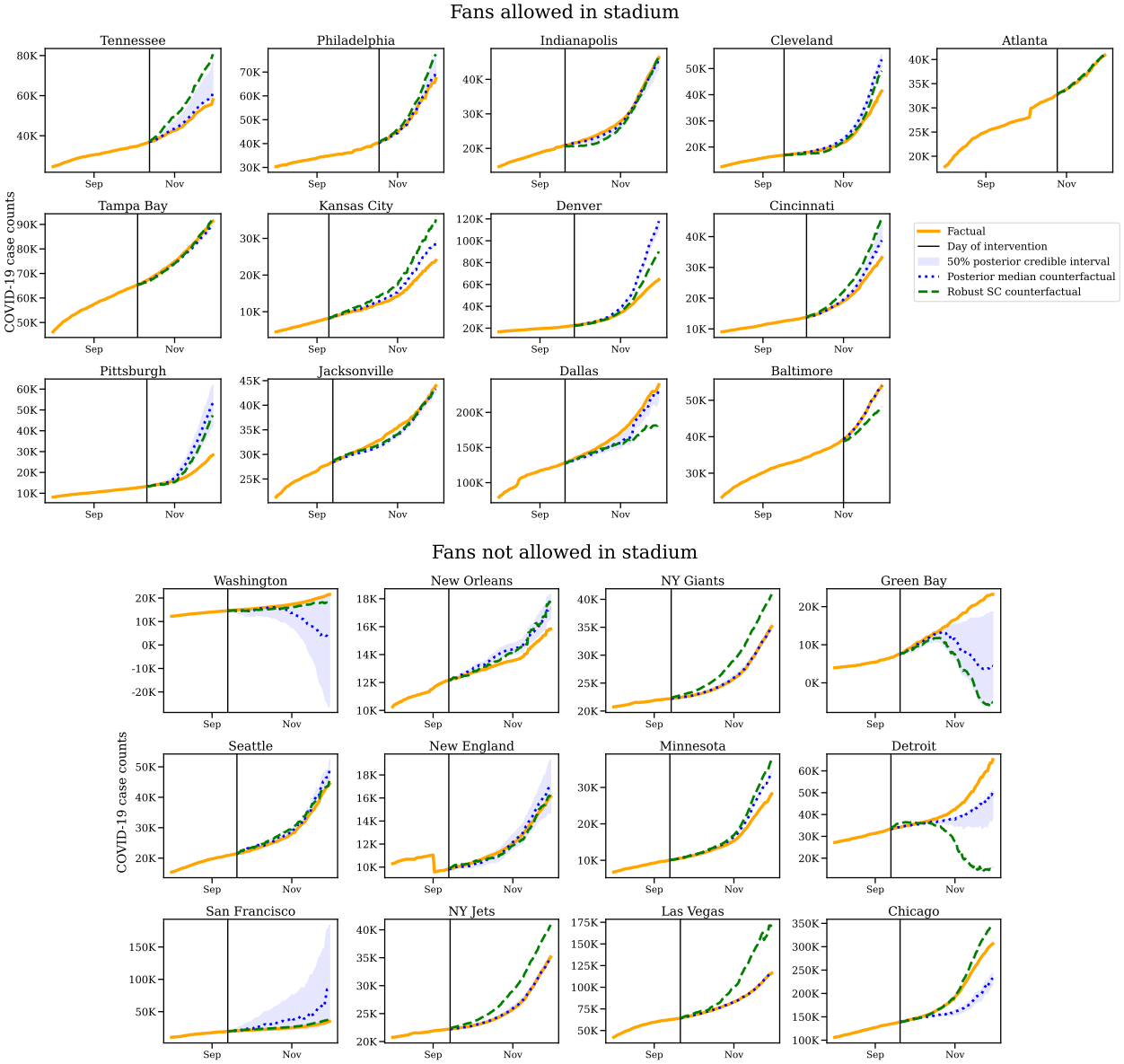


Figure 5: Counterfactual (top) and placebo (bottom) trajectories of COVID-19 case counts in each stadium county after fans were allowed by the NFL back in stadiums. The predictions of our approach are largely convergent with those of robust synthetic control, but exhibit fewer large deviations from the factual trajectories.

At a glance, there is fairly high correspondence between the trajectories coming from robust synthetic control (rSC) approach of García Bulle et al. (2022) and those coming from this paper's approach, as one might expect from the overall correspondence in total treatment effects estimated by both. However there are some noticeable differences between the two. First, by our count, in 13 of the 25 plots, the blue trajectory (this paper) is closer to the orange one (factual) than the green (rSC), while

the green is closer in only five plots; in the remaining seven, blue and green are virtually the same. Since the overall evidence is consistent with a null treatment effect, we surmise that counterfactual trajectories that are closer to the factual ones are likely more accurate, though this is ultimately un-checkable. For the placebo trajectories though, there are three instances where our approach predicts very close factual trajectory, while rSC deviates—specifically, NY Giants, NY Jets, and Las Vegas. In one instance, Chicago, our approach deviates substantially from the factual approach, while rSC does not. We see Green Bay and Detroit, where both deviate but rSC does much more so. Finally, we notice there are other instances where the posterior median trajectory deviates, but the posterior is highly uncertain and its interquartile range covers both the factual and rSC trajectories (e.g., San Francisco). Altogether, the results suggest that proposed approach is broadly consistent with rSC, but possibly more stable in its predictions.

5. Semi-synthetic study: How does the change of stadium policies affect COVID rates?

The previous section's results suggested that the proposed approach's counterfactual predictions, while broadly consistent with rSC, were more stable and less subject to seemingly wild fluctuation. If true, one possibility as to why would be the difference in the probabilistic assumptions of the two approaches. The proposed approach relied on a probabilistic factor model which assumes the data are conditionally Poisson-distributed. This assumption is arguably more appropriate for bursty COVID case counts than the Gaussian assumption implicitly encoded in the linear factor model on which rSC relies. Another possibility as to why is the proposed approach being based on Bayesian inference rather than maximum likelihood like rSC. It is plausible that either the priors serve to regularize the proposed approach's predictions, or posterior averaging serves to smooth them.

In this section, we adopt a semi-synthetic study design to study these questions in more detail. We begin with the real COVID case count data in the previous section. For a randomly-selected subset of the teams—Baltimore, Philadelphia, Minnesota, and New Orleans—we regenerate their (counties $\times$ time steps) count matrix according to a probabilistic factor model that we fit to the original matrix $\mathbf{Y}^{\text{orig}}$. In so doing, we create a semi-synthetic count matrix $\mathbf{Y}^{\text{synth}}$ that is close to the original, but whose generative process conforms to a factor model, and is furthermore entirely known to us in all other aspects. For ease of notation, in this section we drop the (synth) notation, so that $\mathbf{Y} \equiv \mathbf{Y}^{\text{synth}}$.

In more detail, we first fit a K-dimensional latent factor model to the entire original matrix:

$$\widehat{U}, \widehat{V} \leftarrow \underset{U,V}{\arg\max} \left[\prod_i \prod_t \text{Pois}\left(Y^{\text{orig}}_{i,t}; U_i^\top V_t \right) \right], \tag{46}$$

which returns two non-negative factor matrices $\widehat{U} \in \mathbb{R}_+^{K \times n}$ and $\widehat{V} \in \mathbb{R}_+^{K \times m}$. Recall that n is the number of (both donor and target) counties and m is the number of (both pre- and post-intervention) time steps. To create the synthetic data, we first generate the entire matrix of untreated potential outcomes directly from the fitted factor model as

$$Y^{(0)}_{i,t} \sim \text{Pois}\left(\widehat{U}_i^\top \widehat{V}_t \right) \quad \text{for all } i, t. \tag{47}$$

We then generate the treated potential outcomes as

$$Y^{(1)}_{i,t} \sim \text{Pois}\left(\widehat{U}_i^\top \widehat{V}_t \cdot \delta_t \right) \quad \text{for } i \text{ with } A_i = 1 \text{ and } t > 0, \tag{48}$$

where δ_t represents a multiplicative causal effect, which is a function of t that we set to

$$\delta_t = [1 + \log(t)]^{\frac{t\alpha}{1000}}. \tag{49}$$

This is an increasing function in t which is parameterized by $\alpha \in \mathbb{R}$. The case of $\alpha = 0$ corresponds to the absence of any causal effect ($\delta_t = 1$), while positive values of $\alpha > 0$ correspond to positive effects ($\delta_t > 1$) and negative values $\alpha < 0$ correspond to negative effects ($\delta_t < 1$). The impact of different values of α on the trajectories of $Y_{i,t}^{(1)}$ is visualized below in Figure 6.

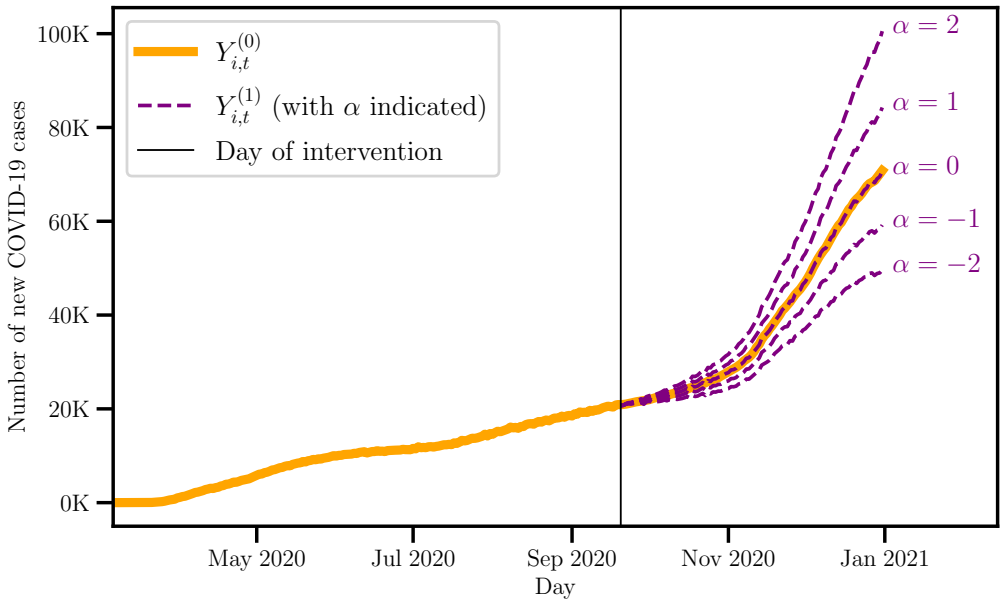


Figure 6: The value of $\alpha \in \mathbb{R}$ scales rate of growth in trajectories of treated potential outcomes, with larger values simulating data from more accelerated epidemics.

Although the conclusions of this study do not hinge on the particular form of δ_t in Equation (49), we took care to select it so that it produced post-intervention trajectories consistent with the geometric growth that characterizes epidemics.

For a given α, we thus generate the treated and untreated potential outcomes and then define the entries of the synthetic data matrix to be $Y_{i,t} = Y_{i,t}^{A_i \cdot \mathbb{I}(t>1)}$, where we keep A_i fixed to the true values.

The benefit of this approach is that we can generate synthetic panel data $\boldsymbol{Y}$ for which we know

1. the true ATT, since we have both sets of potential outcomes, $\boldsymbol{Y}^{(0)}$ and $\boldsymbol{Y}^{(1)}$,
2. that the matrix of untreated potential outcomes $\boldsymbol{Y}^{(0)}$ exactly follows a factor model,
3. the exact form of that factor model—i.e., the rank K, the Poisson likelihood, etc.

With this in place, we can evaluate how well different methods estimate the ATT from $\boldsymbol{Y}$ by comparing their estimates to the ground truth ATT. Moreover, in terms of the different assumptions on which competing methods rely (e.g., Poisson versus Gaussian likelihoods), we can study whether and how much deviations from the known underlying generative process affect performance.

We compare the three methods mentioned in the previous section—i.e.,

1. robust synthetic control (rSC),
2. the proposed approach using probabilistic PCA as the Bayesian factor model (PPCA),
3. the proposed approach using the Gamma-Poisson factor model (GaP)

The first two methods (rSC and PPCA) assume the data are Gaussian, while the last (GaP) assumes they are Poisson. Meanwhile, the latter two methods (PPCA and GaP) are based on Bayesian inference, while the first (rSC) is based on MLE. Systematic differences in performance may stem from differences in probabilistic assumptions or differences in the class of inference procedures—the inclusion of PPCA as a baseline method should thus help us distinguish between these, since it overlaps with rSC on probabilistic assumptions but with GaP on inference class.

All three of these methods ultimately rely on the assumption that the matrix of untreated potential outcomes follows a factor model. To further study the robustness of these methods to departures from that assumption, we also modified the generative process to accommodate differing degrees of factor model mismatch. Specifically, we define the following modification of Equation (47),

$$Y_{i,t}^{(0)} \sim \text{Pois}\left((1-\rho)\,\widehat{U}_i^\top \widehat{V}_t \;+\; \rho e_{i,t} Y_{i,t-1}^{(0)}\right), \tag{50}$$

where $\rho \in [0,1]$ defines a mixture between the factor model and an autoregresive model with random multiplicative innovations $e_{i,t}$. To pronounce the departure of the autoregressive component from the factor model, we define the distribution of the random innovations to be periodic,

$$e_{i,t} \sim 0.9 + 0.2\ \text{Beta}\left(1 + 2\,\mathbb{I}\big(t \bmod 10 < 5\big),\ 1 + 2\,\mathbb{I}(t \bmod 10 \geq 5)\right), \tag{51}$$

which leads to complex and unpredictable trajectories, as depicted below in Figure 7.

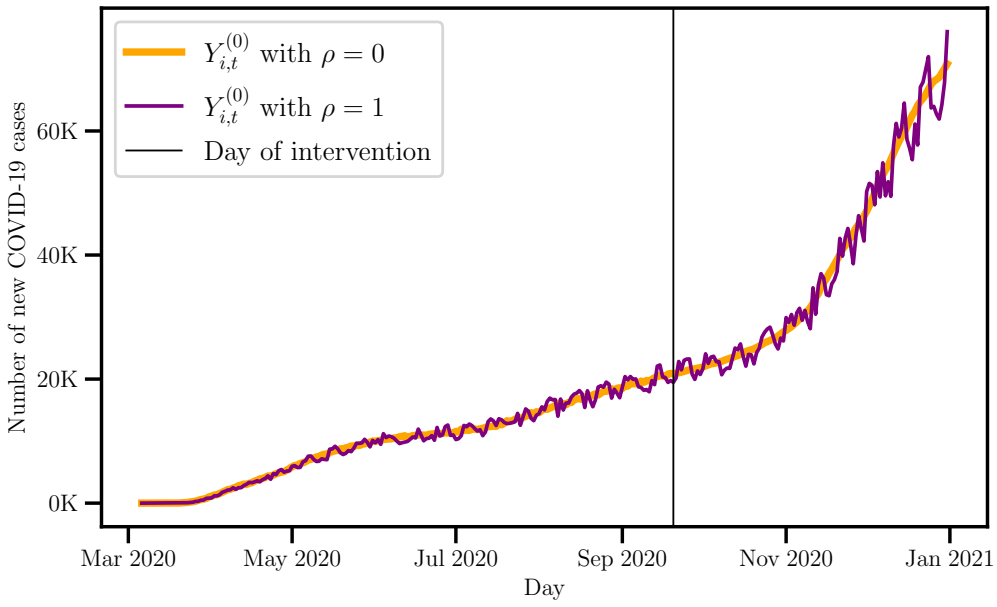


Figure 7: The parameter $\rho \in [0,1]$ is the degree of departure from the factor model assumption.

Finally, in addition to varying true causal effect via α, and the degree of departure from the factor model via ρ, we also vary the number of pre-intervention time steps $\overleftarrow{m}$ that are available for training. Since the proposed approach (GAP and PPCA) relies explicitly on the assumption of many negative control outcomes while rSC does not, it is worth comparing the performance of the two approaches under different settings of small versus large $\overleftarrow{m}$.

We randomly generate four synthetic data sets for every combination of the following values: team $\in$ {Baltimore, Minnesota, New Orleans, Philadelphia, $\rho \in \{0, 1/2, 1\}$, $\alpha \in \{1/10, 4\}$, $\overleftarrow{m} \in \{25, 100\}$, and $K \in \{10\}$. For each combination, we then fit rSC, GAP, and PPCA, using the true K, and otherwise the same settings as described in the previous section.

Each method produces a counterfactual prediction of $\widehat{Y}_{i,t}^{(0)}$ for i with $A_i = 1$ and $t > 0$. In the case of GaP and PPCA, which produce a set of posterior samples, we consider the posterior mean.

We evaluate each method's counterfactual predictions using mean relative error (MRE), which is defined as follows for a given synthetic data set, a given model's prediction, and a given post-intervention time step t:

$$\text{MRE}_t = \frac{1}{\sum_{i=1}^{n} A_i} \sum_{i:A_i=1} \frac{|Y_{i,t}^{(0)} - \widehat{Y}_{i,t}^{(0)}| + \epsilon}{Y_{i,t}^{(0)} + \epsilon}$$

Here $\epsilon > 0$ prevents the denominator from being zero in cases of zero-valued counts. In our setting, where the COVID case counts are safely away from zero, we can use $\epsilon = 0$. Note that in these synthetic data sets, there is only one unit with $A_i = 1$, so the mean above contains only a single element. However, we further average relative error across the four random generations of a given combination of values, as well as the four NFL teams; each value of MRE_t, is thus averaged across 16 values of relative error.

Figure 8 gives the results. Each subplot depicts MRE_t across 30 post-intervention days for each of the three baselines, along with bands reflecting its 95% confidence interval, for one combination of experimental settings. The top six subplots all pertain to the large causal effect of $\alpha = 4$ while the bottom subplots all pertain to the small causal effect of $\alpha = 1/10$. Within each of these two clusters, the first, second, and third rows pertain to $\rho = 0$ (no model mismatch), $\rho = 1/2$ (medium mismatch), and $\rho = 1$ (high mismatch), respectively. Finally, the first and second columns of the subplots then pertain to $\overleftarrow{m} = 25$ and $\overleftarrow{m} = 100$, respectively.

Several patterns are evident. First, GaP, the only model whose probabilistic assumptions match the Poisson likelihood of the true generative process, obtains the lowest MRE_t in virtually all settings. This is even true as the generative process deviates from the factor model for $\rho > 0$. Note that even as we deviate from a factor model, the family of GaP's likelihood (Poisson) is still correct. Indeed, as ρ increases, the performance of GaP appears stable, while the error of the Gaussian models seems to increase.

Second, the error of rSC and PPCA often track each other, suggesting that the (incorrect) Gaussian assumption they share plays a more important role than the style of inference (Bayesian versus MLE) on which they differ.

Third, the performance of all three models is closer when there are fewer observed pre-intervention time steps. When there are more, the performance gap between GaP and the Gaussian models becomes much more pronounced, suggesting that the combination of observing many negative control outcomes and having a well-fitting factor model is necessary for good performance, just as the theory in Section 3 stipulates.

Finally, the results seem stable across the different causal effect sizes α, which does not appear to affect performance.

Altogether, these results strongly suggest that by permitting the use of a much broader range of (possibly non-linear) probabilistic factor models, the proposed approach can strongly outperform rSC, and do so merely by adopting a factor model which assumes a likelihood appropriate for the data. The gap in performance seems to persist even under deviations from the assumption of an underlying factor model, and does not seem attributable solely to the use of Bayesian inference over MLE.

6. Discussion

In this work, we study how the external data of many negative control outcomes can help control for unobserved confounding in causal inference. We develop an algorithm that handles unobserved multi-period confounders with probabilistic factor models. We describe the theoretical assumptions required for the algorithm. We also show how the proposed algorithm generalizes the synthetic control method in settings where the number of units and the number of pre-treatment periods go to infinity simultaneously. This connection justifies the synthetic control method beyond linear models. We demonstrate in simulated and real case studies that both the proposed algorithm and the synthetic control method are robust to certain unobserved confounding.

There are many promising venues for future work. The theoretical results around the proposed algorithm focuses on causal identification with infinite number of control outcomes, i.e. given infinite sample size n and infinite control outcomes m. What are the finite-sample properties of the proposed algorithm? How does the algorithm fare in small-sample regimes, e.g. settings where the synthetic control method is commonly applied? Can we derive partial identification intervals when we have only a limited number of control outcomes? How can we rigorously quantify the uncertainty of the algorithm in these contexts? How can we extend the proposed algorithm to handle outcome of interest that are multi-dimensional or are stochastic processes?

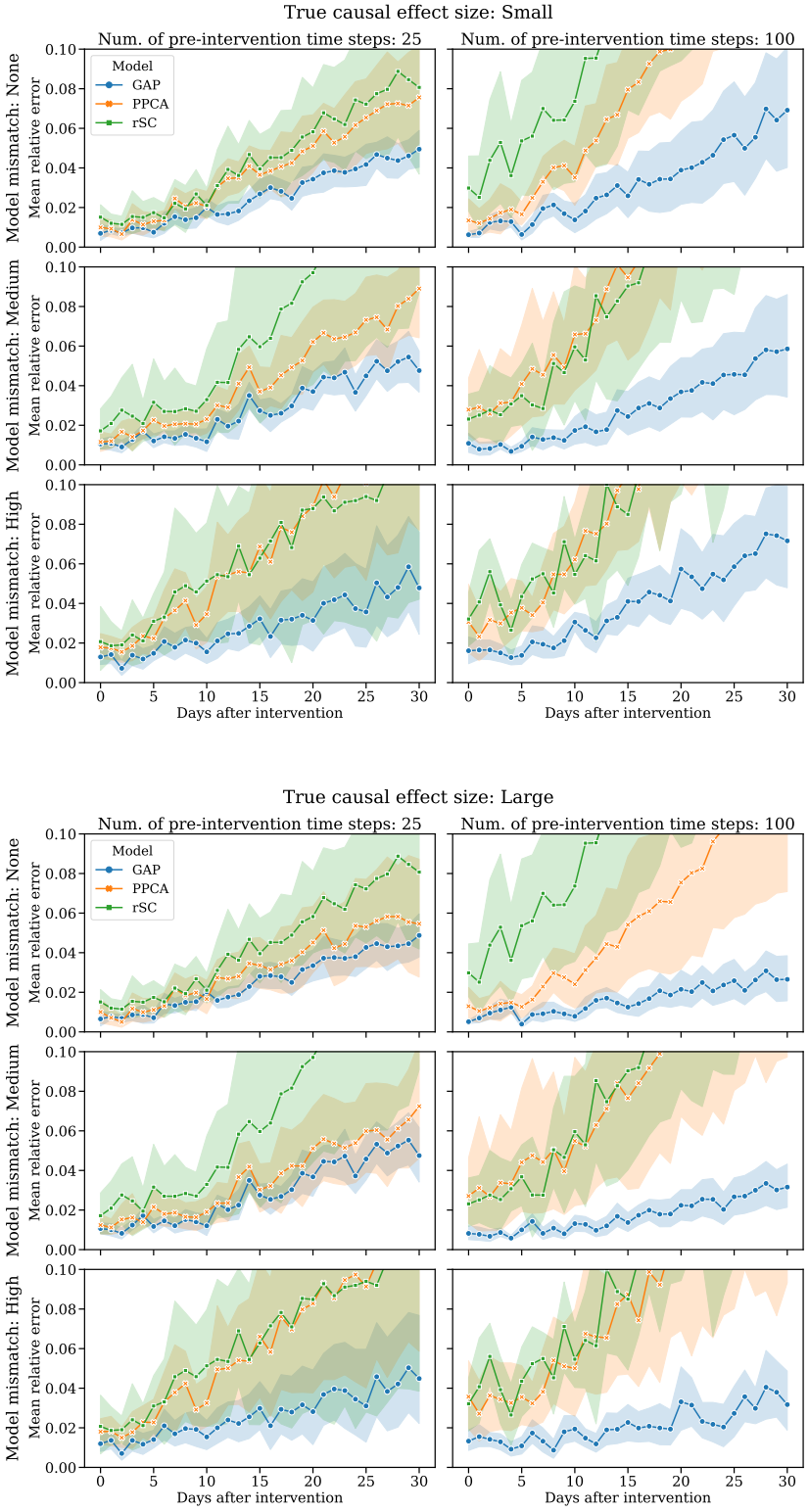


Figure 8: Mean relative error of the counterfactual trajectories predicted by robust synthetic control (rSC) and our approach using PPCA and gamma-Poisson (GAP) as the factor model. The GAP model matches the true data generating process, and the error of its predictions is lowest across all settings. Meanwhile, PPCA and rSC both make Gaussian assumptions about the data, and their error tracks across the range of settings.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1), 1–19.

Abadie, A., Diamond, A., & 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.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2), 495–510.

Abadie, A. & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1), 113–132.

Abadie, A. & L'Hour, J. (2017). A penalized synthetic control estimator for disaggregated data. *Work. Pap., Mass. Inst. Technol., Cambridge, MA*.

Abraham, S. & Sun, L. (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Available at SSRN 3158747*.

Agarwal, A., Dahleh, M., Shah, D., & Shen, D. (2023). Causal matrix completion. In *The Thirty Sixth Annual Conference on Learning Theory* (pp. 3821–3826).: PMLR.

Amjad, M., Misra, V., Shah, D., & Shen, D. (2019). mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2), 37.

Amjad, M., Shah, D., & Shen, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, 19(1), 802–852.

Angrist, J. D. & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2019). *Synthetic difference in differences*. Technical report, National Bureau of Economic Research.

Arnold, B. F., Ercumen, A., Benjamin-Chung, J., & Colford Jr, J. M. (2016). Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 27(5), 637.

Arora, S., Ge, R., et al. (2013). A practical algorithm for topic modeling with provable guarantees. In *International conference on machine learning* (pp. 280–288).: PMLR.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2018). *Matrix completion methods for causal panel data models*. Technical report, National Bureau of Economic Research.

Athey, S. & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2), 431–497.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1), 135–171.

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536), 1789–1803.

Ben-Michael, E., Feller, A., & Rothstein, J. (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2), 351–381.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1), 249–275.

Bing, X., Bunea, F., Ning, Y., & Wegkamp, M. (2020). Adaptive estimation in structured factor models with applications to overlapping clustering.

Bingham, E., Chen, J. P., et al. (2019). Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1), 973–978.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1, 203–232.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.

Bonhomme, S. & Sauder, U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, 93(2), 479–494.

Botosaru, I. & Gutierrez, F. H. (2018). Difference-in-differences when the treatment status is observed in only one period. *Journal of Applied Econometrics*, 33(1), 73–90.

Bottmer, L., Imbens, G. W., Spiess, J., & Warnick, M. (2024). A design-based perspective on synthetic control methods. *Journal of Business & Economic Statistics*, 42(2), 762–773.

Brodersen, K. H., Gallusser, F., et al. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1), 247–274.

Callaway, B., Li, T., & Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2), 395–413.

Callaway, B. & Sant'Anna, P. H. (2019). Difference-in-differences with multiple time periods. *Available at SSRN 3148250*.

Canny, J. (2004). Gap: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 122–129).

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review*, 43(2), 245–257.

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorization models. *Computational Intelligence and Neuroscience*, 2009.

Cengiz, D., Dube, A., Lindner, A., & Zipperer, B. (2019). *The effect of minimum wages on low-wage jobs: Evidence from the United States using a bunching estimator*. Technical report, National Bureau of Economic Research.

Chen, J. (2023). Synthetic control as online linear regression. *Econometrica*, 91(2), 465–491.

Chen, Y., Li, X., & Zhang, S. (2017). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *arXiv preprint arXiv:1712.08966*.

Chen, Y., Li, X., & Zhang, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, 115(532), 1756–1770.

Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2017). An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*.

Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2018). Practical and robust t-test based inference for synthetic control and related methods.

Chiang, Y. A., Pacca, L., et al. (2025). Trajectories and patterns of US counties' policy responses to the COVID-19 pandemic: A sequence analysis approach. *SSM-Population Health*, 29, 101734.

De Chaisemartin, C. & D'HaultfŒuille, X. (2017). Fuzzy differences-in-differences. *The Review of Economic Studies*, 85(2), 999–1028.

Doudchenko, N. & Imbens, G. W. (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis*. Technical report, National Bureau of Economic Research.

Dube, A. & Zipperer, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

Dusetzina, S. B., Brookhart, M. A., & Maciejewski, M. L. (2015). Control outcomes and exposures for improving internal validity of nonrandomized studies. *Health services research*, 50(5), 1432–1451.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Egami, N. (2018). Identification of causal diffusion effects using stationary causal directed acyclic graphs. *arXiv preprint arXiv:1810.07858*.

Ferman, B. (2019). On the properties of the synthetic control estimator with many periods and many controls.

Flanders, W. D., Strickland, M. J., & Klein, M. (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *American journal of epidemiology*, 185(10), 941–949.

Franks, A., D'Amour, A., & Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, (just-accepted), 1–38.

García Bulle, B., Shen, D., Shah, D., & Hosoi, A. E. (2022). Public health implications of opening National Football League stadiums during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, 119(14), e2114226119.

Gilbert, P. B., Bosch, R. J., & Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3), 531–541.

Gobillon, L. & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3), 535–551.

Goodman-Bacon, A. (2018). *Difference-in-differences with variation in treatment timing*. Technical report, National Bureau of Economic Research.

Gopalan, P., Hofman, J. M., & Blei, D. M. (2013). Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*.

Gopalan, P., Hofman, J. M., & Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*.

Gopalan, P. K., Charlin, L., & Blei, D. (2014). Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems* (pp. 3176–3184).

Gunsilius, F. F. (2023). Distributional synthetic controls. *Econometrica*, 91(3), 1105–1117.

Hamad, R., Lyman, K. A., et al. (2022). The US COVID-19 County Policy Database: a novel resource to support pandemic-related research. *BMC Public Health*, 22(1), 1882.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). *Characterizing selection bias using experimental data*. Technical report, National bureau of economic research.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4), 605–654.

Hirano, K. & Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, 226164, 73–84.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1), 1593–1623.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Hsiao, C., Steve Ching, H., & Ki Wan, S. (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5), 705–740.

Imai, K. & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2), 467–490.

Imai, K. & 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.

Imbens, G. & Rubin, D. (2015). *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3), 706–710.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2), 126–132.

Janzing, D. & Schölkopf, B. (2018a). Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1).

Janzing, D. & Schölkopf, B. (2018b). Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810*.

Kasza, J., Wolfe, R., & Schuster, T. (2017). Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *International journal of epidemiology*, 46(4), 1303–1311.

Kinn, D. (2018). Synthetic control methods and big data. *arXiv preprint arXiv:1803.00096*.

Koller, D. & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

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

Li, K. T. (2019). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, (pp. 1–16).

Lipsitch, M., Tchetgen, E. T., & Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3), 383.

Liu, F. & Chan, L. (2018). Confounder detection in high dimensional linear models using first moments of spectral measures. *arXiv preprint arXiv:1803.06852*.

Madigan, D., Stang, P. E., et al. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1, 11–39.

McLachlan, G. J. & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker.

Miao, W., Geng, Z., & Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4), 987–993.

Miao, W. & Tchetgen, E. T. (2018). A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*.

Miao, W. & Tchetgen Tchetgen, E. (2017). Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *American journal of epidemiology*, 185(10), 950–953.

Moran, G. E., Blei, D. M., & Ranganath, R. (2024). Holdout predictive checks for bayesian model criticism. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1), 194–214.

Ogburn, E. L., Shpitser, I., & Tchetgen, E. J. T. (2019). Comment on "blessings of multiple causes". *Journal of the American Statistical Association*, 114(528), 1611–1615.

Ogburn, E. L., Shpitser, I., & Tchetgen, E. J. T. (2020). Counterexamples to" the blessings of multiple causes" by wang and blei. *arXiv preprint arXiv:2001.06555*.

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, 16(1-2), 1–21.

Pearl, J. (2009). *Causality*. Cambridge university press.

Qin, J., Zhang, & Biao (2008). Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2), 329–349.

Ranganath, R., Tang, L., Charlin, L., & Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics* (pp. 762–771).

Richardson, D. B., Keil, A., Tchetgen, T. E., & Cooper, G. (2015). Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology (Cambridge, Mass.)*, 26(5), 727.

Robbins, M. W., Saunders, J., & Kilmer, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *Journal of the American Statistical Association*, 112(517), 109–126.

Robins, J. M., Rotnitzky, A., & Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (pp. 1–94). Springer.

Rosenbaum, P. R. (1989). The role of known effects in observational studies. *Biometrics*, 45(2), 557–569.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4), 472–480.

Samartsidis, P., Seaman, S. R., et al. (2019). Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statistical Science*, 34(3), 486–503.

Sanderson, E., Macdonald-Wallis, C., & Davey Smith, G. (2017). Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International journal of epidemiology*, 47(2), 587–596.

Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., & Suchard, M. A. (2016). Robust empirical calibration of p-values using observational data. *Statistics in medicine*, 35(22), 3883.

Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., & Suchard, M. A. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11), 2571–2577.

Sharma, A., Hofman, J. M., & Watts, D. J. (2016). Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv preprint arXiv:1611.09414*.

Shi, C., Sridhar, D., Misra, V., & Blei, D. (2022). On the assumptions of synthetic control methods. In *International Conference on Artificial Intelligence and Statistics* (pp. 7163–7175).: PMLR.

Shi, X., Miao, W., & Tchetgen, E. T. (2018). Multiply robust causal inference with double negative control adjustment for unmeasured confounding. *arXiv preprint arXiv:1808.04906*.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., & Tchetgen, E. J. T. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3), 348.

Spiess, J., Venugopal, A., et al. (2024). Double and single descent in causal inference with an application to high-dimensional synthetic control. *Advances in Neural Information Processing Systems*, 36.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4), 795–809.

Sun, L., Ben-Michael, E., & Feller, A. (2023). Using multiple outcomes to improve the synthetic control method. *arXiv preprint arXiv:2311.16260*.

Tanaka, M. (2021). Bayesian matrix completion approach to causal inference with panel data. *Journal of Statistical Theory and Practice*, 15(2), 49.

Tchetgen Tchetgen, E. (2013). The control outcome calibration approach for causal inference with unobserved confounding. *American journal of epidemiology*, 179(5), 633–640.

Tipping, M. E. & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611–622.

Viviano, D. & Bradic, J. (2019). Synthetic learner: model-free inference on treatments over time. *arXiv preprint arXiv:1904.01490*.

Wang, Y. & Blei, D. M. (2019a). The blessings of multiple causes. *Journal of the American Statistical Association*, 0(ja), 1–71.

Wang, Y. & Blei, D. M. (2019b). The blessings of multiple causes: Rejoinder. *Journal of the American Statistical Association*, 114(528), 1616–1619.

Wang, Y. & Blei, D. M. (2020). Towards clarifying the theory of the deconfounder. *arXiv:2003.04948*.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1), 57–76.

Zeitler, J., Vlontzos, A., & Gilligan-Lee, C. M. (2023). Non-parametric identifiability and sensitivity analysis of synthetic control models. In *Conference on Causal Learning and Reasoning* (pp. 850–865).: PMLR.

Zheng, J., Wu, J., D'Amour, A., & Franks, A. (2023). Sensitivity to unobserved confounding in studies with factor-structured outcomes. *Journal of the American Statistical Association*, (pp. 1–12).

Zhou, Y., Tang, D., Kong, D., & Wang, L. (2024). Promises of parallel outcomes. *Biometrika*, 111(2), 537–550.

Supplementary Material

Appendix A. Discussion

In this work, we study how the external data of many negative control outcomes can help control for unobserved confounding in causal inference. We develop an algorithm that handles unobserved multi-period confounders with probabilistic factor models. We describe the theoretical assumptions required for the algorithm. We also show how the proposed algorithm generalizes the synthetic control method in settings where the number of units and the number of pre-treatment periods go to infinity simultaneously. This connection justifies the synthetic control method beyond linear models. We demonstrate in simulated and real case studies that both the proposed algorithm and the synthetic control method are robust to certain unobserved confounding.

There are many promising venues for future work. The theoretical results around the proposed algorithm focuses on causal identification with infinite number of control outcomes, i.e. given infinite sample size n and infinite control outcomes m. What are the finite-sample properties of the proposed algorithm? How does the algorithm fare in small-sample regimes, e.g. settings where the synthetic control method is commonly applied? Can we derive partial identification intervals when we have only a limited number of control outcomes? How can we rigorously quantify the uncertainty of the algorithm in these contexts? How can we extend the proposed algorithm to handle outcome of interest that are multi-dimensional or are stochastic processes?

Appendix B. Proof of Theorem 4

Proof Assumptions 1 to 5 guarantee that the unobserved multi-period confounder U_i and the covariates X_i satisfy weak unconfoundedness. In addition, the estimated confounding structure Z_i is consistent with U_i and can be determined in theory from the observed data. Further, Assumptions 4 and 5 ensures that overlap holds with Z_i, X_i. Hence, we can treat Z_i as if it were observed confounders as X_i and proceed with causal inference:

$$\begin{aligned}
\mathbb{E}\left[Y_i(a)\right] &= \mathbb{E}_{Z_i,X_i}\left[\mathbb{E}\left[Y_i(a) \,|\, Z_i, X_i\right]\right] && (52)\\
&= \mathbb{E}_{Z_i,X_i}\left[\mathbb{E}\left[Y_i(a) \,|\, A_i = a, Z_i, X_i\right]\right] && (53)\\
&= \mathbb{E}_{Z_i,X_i}\left[\mathbb{E}\left[Y_i \,|\, A_i = a, Z_i, X_i\right]\right], && (54)
\end{aligned}$$

where the first equality is due to the tower property, the second equality is due to Assumption 3, and the third equality is due to SUTVA.

Similarly, we have

$$\begin{aligned}
\mathbb{E}\left[Y_i(a) \,|\, A_i = a'\right] &= \mathbb{E}_{Z_i,X_i \,|\, A_i=a'}\left[\mathbb{E}\left[Y_i(a) \,|\, A_i = a', Z_i, X_i\right]\right] && (55)\\
&= \mathbb{E}_{Z_i,X_i \,|\, A_i=a'}\left[\mathbb{E}\left[Y_i(a) \,|\, A_i = a, Z_i, X_i\right]\right] && (56)\\
&= \mathbb{E}_{Z_i,X_i \,|\, A_i=a'}\left[\mathbb{E}\left[Y_i \,|\, A_i = a, Z_i, X_i\right]\right]. && (57)
\end{aligned}$$

■

Appendix C. Proof of ??

We first prove the first part.

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} f_y(Y_{i,t}^{(0)})\right] \tag{58}$$

$$=\mathbb{E}\left[\frac{1}{nP(A_i=1)}\sum_{i=1}^{n}\mathbb{I}\{A_i=0\}\frac{P(A_i=1\,|\,Z_i,X_i)}{P(A_i=0\,|\,Z_i,X_i)}f_y(Y_{i,t}^{(0)})\right] \tag{59}$$

$$=\frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n}\frac{P(A_i=1\,|\,Z_i,X_i)}{P(A_i=0\,|\,Z_i,X_i)}\mathbb{E}\left[\mathbb{I}\{A_i=0\}f_y(Y_{i,t}^{(0)})\,|\,Z_i,X_i\right]\right] \tag{60}$$

$$=\frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n}\frac{P(A_i=1\,|\,Z_i,X_i)}{P(A_i=0\,|\,Z_i,X_i)}\mathbb{E}\left[\mathbb{I}\{A_i=0\}\,|\,Z_i,X_i\right]\mathbb{E}\left[f_y(Y_{i,t}^{(0)})\,|\,Z_i,X_i\right]\right] \tag{61}$$

$$=\frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{I}\{A_i=1\}\,|\,Z_i,X_i)\right]\mathbb{E}\left[f_y(Y_{i,t}^{(0)})\,|\,Z_i,X_i\right]\right] \tag{62}$$

$$=\frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{I}\{A_i=1\}f_y(Y_{i,t}^{(0)})\,|\,Z_i,X_i)\right]\right] \tag{63}$$

$$=\frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}\{A_i=1\}f_y(Y_{i,t}^{(0)})\right] \tag{64}$$

$$=\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n}\mathbb{I}\{A_i=1\}}\sum_{A_i=1}f_y(Y_{i,t}^{(0)})\right] \tag{65}$$

The first equality is due to the definition of the inverse probability weights. The second equality is due to the tower property. The third equality is due to $f_y(Y_{i,t}^{(0)}) \perp A_i \,|\, Z_i, X_i$. The fourth equality is due to canceling out $P(A_i = 0 \,|\, Z_i, X_i)$. The fifth equality is again due to $f_y(Y_{i,t}^{(0)}) \perp A_i \,|\, Z_i, X_i$. The sixth equality is due to the tower property.

We can prove the other equalities of the first part with the exact same argument.

We next prove the second part.

First, when $m \to \infty$ (due to Assumption 2),

$$P(Z_i \,|\, Y_{i,t}, \{Y_{i,t}^{(0)}\}_t) = \delta_{f(Y_i, \{Y_{i,t}^{(0)}\}_t)} = \delta_{\tilde{f}(\{Y_{i,t}^{(0)}\}_t)}, \tag{66}$$

for some function $\tilde{f}$. The reason is that when $m \to \infty$, any infinite subset of $Y_{i,t}, \{Y_{i,t}^{(0)}\}_t$ would determine Z_i (Chen et al., 2017). Hence $\{Y_{i,t}^{(0)}\}_t$ would be sufficient to determine Z_i in thoery, i.e. $Z_i \perp Y_{i,t} \,|\, \{Y_{i,t}^{(0)}\}_t$. Therefore, Equation (66) holds and $\tilde{f}$ must be linear as is f, which implies

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} Z_i\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n}\mathbb{I}\{A_i=1\}}\sum_{A_i=1} Z_i\right], \tag{67}$$

given **??**.

Next, using the Lagrangian multipliers, one can show the solution to the optimization problem with constraints **????** and eq. (67) is

$$w_i^{\text{SC}} = \exp(\alpha + \theta_Z Z_i + \theta_X \cdot X_i)), \tag{68}$$

where $\alpha = \log \frac{P(A_i=1)}{P(A_i=0)}$ (Cover & Thomas, 2012; Zhao & Percival, 2017). This solution of w_i^{SC} coincides with w_i^{IPW} when the propensity score model is log-linear in Z_i, X_i. Moreover, w_i^{SC} satisfies **??** because $Y_{i,t}^{(0)}$ is weak unconfounded given Z_i, X_i (Assumptions 3 and 4).

References

Chen, Y., Li, X., & Zhang, S. (2017). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *arXiv preprint arXiv:1712.08966*.

Cover, T. M. & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Zhao, Q. & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1).