Structural Nested Mean Models Under Parallel Trends Assumptions

Zach Shahn[1,2], Oliver Dukes[3], David Richardson[4], Eric Tchetgen Tchetgen[3], and James Robins[5]

[1]CUNY School of Public Health, New York, NY, USA
[2]IBM Research, Yorktown Heights, NY, USA
[3]University of Pennsylvania, Philadelphia, PA, USA
[4]University of California Irvine, Irvine, CA, USA
[5]Harvard TH Chan School of Public Health, Boston, MA, USA

May, 2022

Abstract

In this paper, we generalize methods in the Difference in Differences (DiD) literature by showing that both additive and multiplicative standard and coarse Structural Nested Mean Models (Robins, 1994, 1997, 1998, 2000, 2004; Lok and Degruttola, 2012; Vansteelandt and Joffe, 2014) are identified under parallel trends assumptions. Our methodology enables adjustment for time-varying covariates, identification of effect heterogeneity as a function of time-varying covariates, identification of additional causal contrasts (such as effects of a 'blip' of treatment at a single time point followed by no further treatment and controlled direct effects), and estimation of treatment effects under a general class of treatment patterns (e.g. we do not restrict to the 'staggered adoption' setting, and treatments can be multidimensional with any mix of categorical and continuous components). We stress that these extensions come essentially for free, as our parallel trends assumption is not stronger than other parallel trends assumptions in the DiD literature. We also provide a method for sensitivity analysis to violations of our parallel trends assumption. We also explain how to estimate optimal treatment regimes via optimal regime Structural Nested Mean Models under parallel trends assumptions plus an assumption that there is no effect modification by unobserved confounders.

Finally, we illustrate our methods with real data applications estimating effects of bank deregulation on housing prices and effects of floods on flood insurance take-up.

1 Introduction

In this paper, we link and extend two approaches to estimating time-varying treatment effects on repeated continuous outcomes—time-varying Difference in Differences (DiD; see Roth et al (2022) for a review) and Structural Nested Mean Models (SNMMs; see Vansteelandt and Joffe (2014) for a review). In particular, we show that SNMMs, which were previously only known to be nonparametrically identified under a no unobserved confounding assumption, are also identified under a generalized version of the parallel trends assumption typically used to justify time-varying DiD methods. Because SNMMs model a broader set of causal estimands, our results allow practitioners of existing time-varying DiD approaches to address additional types of substantive questions (such as characterization of time-varying effect heterogeneity, estimation of the lasting effects of a "blip" of treatment at a single time point, and others) under similar assumptions. Our results also allow analysts who apply SNMMs under the no unobserved confounding assumption to estimate the same causal effects under alternative identifying conditions and thus potentially to triangulate evidence.

The Difference in Differences design (Snow, 1855; Card and Krueger, 1993) is a popular approach to estimating causal effects in the possible presence of unobserved confounding. The key assumption it requires is 'parallel trends', i.e. the average trend in counterfactual untreated outcomes is equal among the treated and the untreated. The canonical DiD design identifies effects of treatments delivered at a single time period, but recently there has been a significant literature (e.g. Roth et al., 2022; Callaway and Sant'Anna, 2021; Chaisemartin and D'Haultfoeuille, 2021; Athey and Imbens, 2018; Bojinov et al, 2020) on extensions to time-varying treatment strategies.

The time-varying DiD work emphasizes estimating effects of so-called 'staggered adoption' strategies under time-varying versions of the parallel trends assumption, possibly conditional on baseline covariates. In the staggered adoption setting, all units start off at a common baseline level of treatment and then deviate from that baseline level at different times in a staggered fashion. In the simplest and most commonly considered case, the baseline treatment level is 0 or 'untreated', and units then initiate a binary treatment at different times. Callway and Sant'anna (2021) focus on the setting where once treatment is initiated it is sustained until end of follow up. Chaisemartin and D'Haultfoeuille (2021) focus on the setting where after treatment is initiated it may subsequently turn on and off according to the observational treatment regime, and interest centers on intention-to-treat type effects of first initiating treatment. For times $k > m$, time-varying DiD methods identify the effect of starting treatment at time m compared to never starting treatment on the outcome at time k among those who actually started treatment at time m. Different authors propose different weighted averages of these effects over m and k as suitable summaries. Effects conditional on baseline covariates can be identified via subgroup analysis. Applications of time-varying DiD methods often estimate effects of new laws or policies that are implemented in different geographic areas at different times. For example, a researcher might use such an approach to estimate the effect of minimum wage hikes on unemployment rate at the county level.

SNMMs can also be used to estimate time-varying treatment effects. We distinguish between two types of SNMMs—coarse SNMMs (Robins, 1998; Lok and DeGruttola, 2012) and standard SNMMs (Robins, 1994, 1997, 2000, 2004; Vansteelandt and Joffe, 2014). While standard SNMMs were introduced first and are in wider use, coarse SNMMs correspond more closely to staggered adoption strategies. A coarse SNMM models the conditional effect of initiating treatment for the first time at time m compared to never initiating treatment given covariate history through m in those who actually initiated treatment at time m. In settings where subjects switch off and on treatment, coarse SNMMs model conditional intention-to-treat effects of initial departures from the baseline treatment level marginalized over future treatment patterns under the observational regime, like the estimands of Chaisemartin and D'Haultfoeuille (2021) (except conditional on time-varying covariates). In settings where subjects always continue treatment at the same level once initiated, coarse SNMMs model conditional versions of the estimands of Callaway and Sant'anna (2021). Thus, coarse SNMMs are models of time-varying effect heterogeneity of staggered adoption strategies (which could not be characterized by existing DiD approaches). Coarse SNMMs also straightforwardly admit multiplicative effects and multidimensional treatments with continuous and discrete components. For example, a coarse SNMM can model how the effect of a minimum wage hike on county level unemployment rate varies with the previous minimum wage, the county level poverty rate at the time of the wage increase, and the magnitude of the wage increase. The marginal estimands of the time-varying DiD approaches (i.e. not conditional on time-varying covariates), including expected counterfactual outcomes in the full population had treatment never been initiated by anyone, can also be identified in terms of the parameters of a coarse SNMM. Thus, coarse SNMMs estimate a broader class of quantities pertaining to staggered adoption strategies than existing DiD methods. (We also discuss how coarse SNMMs can be used to estimate the controlled direct effect (Robins and Greenland, 1992) of initiating one intervention while keeping another indefinitely at its baseline rate, e.g. the effect of expanding Medicaid and holding minimum wage constant.)

A standard SNMM is a model for the conditional effect of one last 'blip' of treatment delivered at time m followed by no treatment thereafter compared to no treatment from time m onward conditional on previous treatment and covariate history. Note that this blip effect is distinct from the conditional effect of treatment initiation followed by the observational regime (which in some cases may be sustained treatment). The lasting effects of a single blip of exposure (not necessarily the initial exposure) might be of interest in many contexts. For example, in education, a researcher might want to estimate the (conditional or marginal) lasting impact of an intervention (e.g. on class size or teacher training) in a single grade on subsequent student outcomes. Or, in occupational health, the (conditional or marginal) lasting effect of a short exposure to radiation in those exposed might be of interest. Like coarse SNMMs, standard SNMMs are powerful models for effect heterogeneity as a function of time-varying covariates, and many counterfactual quantities and causal contrasts beyond conditional effects of blips of treatment on the treated described above can be identified in terms of their parameters (including, as we show in Section 8, controlled direct effects). Like coarse SNMMs, standard SNMMs can also straightforwardly model multiplicative effects and admit multidimensional treatments with continuous and discrete components.

Past work has shown that coarse and standard SNMM parameters can be consistently estimated via 'g-estimation' under no unobserved confounding assumptions (Robins, 1994, 1997, 1998, 2004). Here, we show that SNMM parameters also can be consistently estimated without a no unobserved confounding assumption. We exchange the assumption of no unmeasured confounders for the alternative assumption of time-varying conditional parallel trends: conditional on observed history through time m, untreated counterfactual outcome trajectories from time m onwards are parallel in the treated and untreated for all m. While neither the no unobserved confounding assumption nor the conditional parallel trends assumption is empirically verifiable, access to alternative approaches that yield identification of causal effects under different identifying conditions can help to triangulate evidence.

Our parallel trends assumption is not stronger than other parallel trends assumptions in the time-varying DiD literature. In fact, our conditional time-varying parallel trends assumption might be considered more plausible than an assumption that conditions only on baseline covariates. For example, if treatment and trends in counterfactual outcomes are both associated with a time-varying covariate, then the baseline version of the parallel trends assumption would not hold, but parallel trends conditional on that time-varying covariate might hold. Thus, our parallel trends assumption is a generalization that allows identification in some additional settings compared to previous time-varying DiD methods.

We also show that if one is willing to assume parallel trends under any given treatment strategy g (as opposed to under sustained baseline treatment), it is possible to identify the parameters of a general regime SNMM (Robins, 2004) and thus the expected outcome trajectory had everybody in the population followed strategy g. g can even be a dynamic regime, i.e. treatment under g can depend on covariate history. Renson et al. independently showed (in a preprint that originally appeared very shortly after and essentially concurrently with the first arxiv version of this paper) that these expectations are identified. We further show that under the much stronger assumptions of parallel trends under all treatment strategies and no effect modification by unobserved confounders (where the second assumption is practically implied by the first, as we demonstrate in the Appendix), it is possible to identify the parameters of an optimal regime SNMM (Robins, 2004) and thus estimate optimal dynamic treatment strategies. This last result might be considered a contribution to the reinforcement learning literature, where optimal regime SNMMs are sometimes deemed 'A-learning' (Schulte et al, 2014).

We note that throughout we are assuming availability of panel data, i.e. data on the same units over time, with repeated outcome measures. Some DiD approaches allow for repeated cross-sectional data (i.e. different units measured at each time point), but our method currently does not. (Of course, units might be geographic and their outcomes composites of individual residents, such as unemployment rate in a state or county. In this case, our method would not require that the same individual residents are surveyed at each time point to estimate the geographic unit level outcome.) Further, many SNMM studies consider settings with an outcome measured only once at the end of follow up, and the methods in this paper would not be applicable to these studies.

The organization of the paper is as follows. In Section 2, we establish notation, assumptions, and review coarse and standard SNMMs. In Section 3, we identify and provide doubly robust estimators for the parameters of a coarse SNMM under our parallel trends assumption. In Section 4, we identify and provide doubly robust estimators of standard SNMM parameters under our parallel trends assumption. (Sections 3 and 4 are structurally and symbolically extremely similar and may even appear repetetive. Nevertheless, we include both sections as they pertain to substantively different model classes.) In Section 5, we consider SNMMs for multiplicative effects. In Section 6, we explain how to estimate the parameters of general regime SNMMs under a parallel trends assumption for the regime of interest and how to estimate the optimal regime via an optimal regime SNMM under parallel trends assumptions for all regimes plus an additional no effect modification by unobserved confounders assumption. In Section 7, we present two applications to real data, a coarse SNMM modeling effects of bank deregulation on housing prices and a standard SNMM modeling effects of floods on flood insurance take-up. In Section 8, we discuss some extensions, including descriptions of how to estimate controlled direct effects and perform sensitivity analysis for violations of conditional parallel trends. The theme of the paper is that it is fruitful to use structural nested mean models under DiD type assumptions.

2 Notation, Assumptions, and SNMMs

Suppose we observe a cohort of N subjects indexed by $i \in \{1, \ldots, N\}$. Assume that each subject is observed at regular intervals from baseline time 0 through end of followup time K, and there is no loss to follow-up. At each time point m, the data are collected on $\mathbb{O}_m = (Z_m, Y_m, A_m)$ in that temporal order. A_m denotes the (possibly multidimensional with discrete and/or continuous components) treatment received at time m, Y_m denotes the outcome of interest at time m, and Z_m denotes a vector of covariates at time m excluding Y_m. Hence Z_0 constitutes the vector of baseline covariates other than Y_0. For arbitrary time varying variable X: we denote by $\bar{X}_m = (X_0, \ldots, X_m)$ the history of X through time m; we denote by $\underline{X}_m = (X_m, \ldots, X_K)$ the future of X from time m through time K; and whenever the negative index X_{-1} appears it denotes the null value with probability 1. Define $\bar{L}_m$ to be $(\bar{Z}_m, \bar{Y}_{m-1}, \bar{A}_{m-1})$, the history through time m excluding (Y_m, A_m). Hence L_0 is Z_0. We adopt the counterfactual framework for time-varying treatments (Robins, 1986) which posits that corresponding to each time-varying treatment regime $\bar{a}_m$, each subject has a counterfactual or potential outcome $Y_{m+1}(\bar{a}_m)$ that would have been observed had that subject received treatment regime $\bar{a}_m$.

Throughout, we make the assumption

$$\textbf{Consistency: } Y_m(\bar{A}_{m-1}) = Y_m \; \forall m \leq K \tag{1}$$

stating that observed outcomes are equal to counterfactual outcomes corresponding to observed treatments. This assumption is necessary to link observed data to the counterfactual data.

We will use slightly different notation and make different assumptions for coarse and standard SNMMs. Coarse SNMMs specifically model effects of initial deviations from a baseline treatment level. Standard SNMMs model the effect of a final blip of treatment at occasion m followed by the baseline level of treatment thereafter. The staggered adoption setting where treatment is sustained once initiated is an important setting in the DiD literature where coarse SNMMs are required. In other settings, both coarse and standard SNMMs might be used. We will describe notation and assumptions for coarse and standard SNMMs in separate subsections below. Throughout, we code treatment such that at time m, 0 denotes the baseline level of treatment at at that time.

2.1 Standard SNMMs

Whenever discussing standard SNMMs, we will assume

$$\textbf{Positivity: } f_{A_m|\bar{L}_m,\bar{A}_{m-1}}(0|\bar{l}_m, \bar{a}_{m-1}) > 0 \text{ whenever } f_{\bar{L}_m,\bar{A}_{m-1}}(\bar{l}_m, \bar{a}_{m-1}) > 0. \quad (2)$$

However, we note that under parametric versions of the models we consider positivity is not strictly necessary. Note that this positivity assumption excludes the staggered adoption setting where $f_{A_m|\bar{L}_m,\bar{A}_{m-1}}(0|\bar{l}_m, \bar{a}_{m-1}) = 0$ if $a_{m-1} \neq 0$.

Define the causal contrasts

$$\gamma^*_{mk}(\bar{l}_m, \bar{a}_m) \equiv E[Y_k(\bar{a}_m, \underline{0}) - Y_k(\bar{a}_{m-1}, \underline{0})|\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m] \quad (3)$$

for all $k > m$. $\gamma^*_{mk}(\bar{a}_m, \bar{l}_m)$ is the average effect at time k among patients with history $(\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m)$ of receiving treatment a_m at time m and then 0 thereafter compared to receiving treatment 0 at time m and thereafter. (Here '0' can be replaced by any baseline value a^*_m.) These contrasts are sometimes called 'blip functions' because they are the effects of one last blip of treatment. Note $\gamma^*_{mk}(\bar{l}_m, \bar{a}_m) = 0$ if $\mathrm{a}_m = 0$.

A parametric standard SNMM imposes functional forms on the blip functions $\gamma^*_{mk}(\bar{A}_m, \bar{L}_m)$ for each $k > m$, i.e.

$$\gamma^*_{mk}(\bar{a}_m, \bar{l}_m) = \gamma_{mk}(\bar{a}_m, \bar{l}_m; \psi^*), \quad (4)$$

where ψ^* is an unknown finite dimensional parameter vector and $\gamma_{mk}(\bar{a}_m, \bar{l}_m; \psi)$ is a known function equal to 0 whenever $\psi = 0$ or $a_m = 0$.

From a policy perspective, our motivation to fit a standard SNMM is typically that were the $\gamma^*_{mk}(\bar{a}_m, \bar{l}_m)$ (i.e. ψ^* under (4)) identified, then for any subject history $(\bar{L}_m = \bar{l}_m, \bar{A}_m = \bar{a}_{m-1})$ of interest, the expected conditional counterfactual outcome trajectory under no further treatment, i.e. $E[\underline{Y}_m(a_{m-1}, \underline{0})|\bar{L}_m = \bar{l}_m, \bar{A}_{m-1} = \bar{a}_{m-1}]$, would also be identified. We could also identify quantities that further condition on treatment at m, i.e. $E[\underline{Y}_m(a_{m-1}, \underline{0})|\bar{L}_m = \bar{l}_m, \bar{A}_m = \bar{a}_m]$. Or we could marginalize over $\bar{L}_m$ to identify $E[\underline{Y}_m(a_{m-1}, \underline{0})|\bar{A}_{m-1} = \bar{a}_{m-1}]$ and $E[\underline{Y}_m(a_{m-1}, \underline{0})|\bar{A}_m = \bar{a}_m]$. The derived quantity $E[\bar{Y}_K(\bar{0})]$, i.e. the expected counterfactual trajectory under no treatment, would also be identified. From a scientific perspective, the goal might be to determine the subset of $(\bar{L}_m, \bar{A}_{m-1})$ that are time dependent effect modifiers of the causal contrast (3).

To illustrate the power of SNMMs, we describe some insights that might be gleaned from an SNMM of nonadherence effects in an arm of a hypothetical RCT, a common SNMM application area (Robins, 1998). Suppose a trial is comparing anti-hypertensive medications, and subjects attend K monthly visits where their blood pressure (the outcome of interest, Y_t at week t) and other measurements L_t such as self-reported stress are recorded. At each visit, it is also ascertained whether subjects were adherent in the previous month (A_{t-1}, 0 indicating adherence and 1 non-adherence). In this setting, interest might primarily center around the quantity $E[\bar{Y}_K(\bar{0})]$, i.e. the expected outcome under full adherence in the treatment arm of interest, which is identified from the blip function and often a target of inference in DiD settings. Knowledge of the blip function (3) would further enable identification of additional quantities. For example, querying the blip function (3) directly yields conditional lasting effects of a brief period of nonadherence given past covariate and adherence history. Perhaps a single initial week of nonadherence generally has major lasting effects, even if followed by perfect adherence thereafter, in which case it is possibly worthwhile to devote significant resources to preventing patients from ever lapsing. Alternatively, perhaps the effect of a first week of nonadherence is generally negligible, and nonadherence effects only begin to accrue after multiple weeks of missed treatment, in which case a more cost effective course of action might be to monitor patients and intervene to encourage adherence only once they begin to lapse. SNMMs can distinguish between these scenarios and inform decision making accordingly. As a final example, it might be of both policy and scientific interest to know whether effects of a blip of nonadherence are significantly stronger in subjects who had high stress levels immediately prior, which can also be inferred from an SNMM. While not nearly exhaustive, we hope that this sample of inferences enabled by SNMMs in a hypothetical setting inspires some enthusiasm for the potential of SNMMs in readers more familiar with the DiD literature, where many of these inferences are not possible. Vansteelandt and Joffe (2014) provide a broader overview of the capabilities of SNMMs.

Robins (1994, 1997, 2004) has shown that $\gamma^* = (\gamma^*_{01}, \ldots, \gamma^*_{(K-1)K})$, the vector of all functions $\gamma^*_{mk} = \gamma^*_{mk}(\bar{a}_m, \bar{l}_m)$ with $m < k \leq K$, is nonparametrically identified and how to consistently estimate the parameter ψ^* of a parametric SNMM under the assumption that there are no unobserved confounders, i.e.

$$Y_k(\bar{a}_{m-1}, \underline{0}) \perp\!\!\!\perp A_m | \bar{A}_{m-1} = \bar{a}_{m-1}, \bar{L}_m \ \forall k > m, \bar{a}_{m-1}.$$

In this paper, we will instead make the parallel trends assumption

Time-Varying Conditional Parallel Trends:

$$\begin{aligned} &E[Y_k(\bar{a}_{m-1}, \underline{0}) - Y_{k-1}(\bar{a}_{m-1}, \underline{0}) | \bar{A}_m = \bar{a}_m, \bar{L}_m] = \\ &E[Y_k(\bar{a}_{m-1}, \underline{0}) - Y_{k-1}(\bar{a}_{m-1}, \underline{0}) | \bar{A}_m = (\bar{a}_{m-1}, 0), \bar{L}_m] \\ &\forall k > m. \end{aligned} \quad (5)$$

When $k = m + 1$, this assumption reduces to

$$E[Y_{m+1}(\bar{a}_{m-1}, 0) - Y_m | \bar{A}_m = \bar{a}_m, \bar{L}_m] = E[Y_{m+1}(\bar{a}_{m-1}, 0) - Y_m | \bar{A}_m = (\bar{a}_{m-1}, 0), \bar{L}_m]. \tag{6}$$

The time-varying conditional parallel trends assumption states that, conditional on observed covariate history through time m and treatment history through time $m - 1$, the expected counterfactual outcome trends setting treatment to 0 from time m onwards do not depend on the treatment actually received at time m. (5) will often be a more plausible assumption than a version that does not condition on time-varying covariates. If $\bar{L}_m$ is associated with both A_m and counterfactual outcome trends, then (5) may hold but an assumption that does not condition on $\bar{L}_m$ would not.

Remark 1. *Note that we have defined $\bar{L}_m$ to not include Y_m, since if it did then when $k = m + 1$ the parallel trends assumption would imply that there is no unobserved confounding, which we do not wish to assume as then the estimators we introduce would not be needed. This means we cannot adjust for the most recent outcome or estimate effect heterogeneity conditional on the most recent outcome with the methods in this paper. However, $\bar{Y}_{m-1}$ is included in $\bar{L}_m$.*

2.2 Coarse SNMMs

If we are only interested in the effects of first treatment initiation (where we use the term 'treatment initiation' to refer to the first departure from the baseline treatment level of 0), then let $T \in \{0, \ldots, K, \infty\}$ denote time of treatment initiation with ∞ denoting the strategy of never initiating treatment. Let $Y_k(m, a_m)$ denote the counterfactual outcome at time $k > m$ under the strategy that initiates treatment at value a_m at time m, with $Y(\infty)$ denoting the counterfactual outcome at time k under the strategy that never initiates treatment. Note $Y_k(\infty) = Y_k(m, a_m)$ for $k \leq m$. An important setting where treatment initiation effects are of sole interest is the staggered adoption setting where $A_m = a_m$ implies that $A_k = a_m$ for all $k \geq m$. However, coarse SNMMs might still be useful even if arbitrary treatment patterns are found in the data.

For coarse SNMM identification, we make a weaker positivity assumption that only holds when treatment has not yet been initiated.

Coarse Positivity: $f_{A_m|\bar{L}_m, T \geq m}(a_m | \bar{l}_m, T \geq m) > 0$ whenever $f_{\bar{L}_m, \bar{A}_{m-1}}(\bar{l}_m, T \geq m) > 0$. (7)

Interest centers on the causal contrasts

$$\gamma^{c*}_{mk}(\bar{l}_m, a_m) \equiv E[Y_k(m, a_m) - Y_k(\infty) | T = m, A_m = a_m, \bar{L}_m = \bar{l}_m] \tag{8}$$

for all $k > m$. These contrasts are conditional treatment effects on the treated of initial departures from baseline treatment levels.

A parametric coarse SNMM imposes functional forms on $\gamma^{c*}_{mk}(\bar{l}_m, a_m)$ for each $k \geq m$, i.e.

$$\gamma^{c*}_{mk}(\bar{l}_m, a_m) = \gamma^{c}_{mk}(\bar{l}_m, a_m; \psi^*_c) \tag{9}$$

where ψ^*_c is an unknown finite dimensional parameter vector and $\gamma^{c}_{mk}(\bar{l}_m, a_m; \psi_c)$ is a known function equal to 0 whenever $\psi_c = 0$.

As in the case of standard SNMMs, additional derived quantities of interest become identified along with the $\gamma^{c*}_{mk}(\bar{l}_m, a_m)$ (i.e. ψ^*_c under (9)). For instance, the expected counterfactual trajectory under no treatment ($E[\bar{Y}_K(\infty)]$) and the expected counterfactual trajectory under no treatment in the treated ($E[\underline{Y}_{m+1}(\infty)|T = m, A_m = a_m]$) would also be identified.

Revisiting the illustrative hypothetical nonadherence example from the previous subsection, if subjects never restart treatment once they stop, then it is only possible to fit a coarse SNMM to the data. Whether or not subjects switch back and forth between adherence and nonadherence, however, a coarse SNMM of the effect of initially skipping treatment (marginalizing over future adherence patterns) enables identification of the primary quantity of interest–$E[\bar{Y}_K(\infty)]$, or the expected outcome under full adherence. Further, (8) tells us how effects of (initial) nonadherence vary with time-varying covariates such as stress at the time of (initial) nonadherence. Knowledge of time-varying effect heterogeneity, which is unavailable from DiD estimators, can both inform scientific understanding and resource allocation for preventing nonadherence.

To identify $\gamma^{c*} = (\gamma^{c*}_{01}, \ldots, \gamma^{c*}_{(K-1)K})$, the vector of all functions $\gamma^{c*}_{mk}(\bar{l}_m, a_m)$ with $m < k \leq K$, we will make the coarse time-varying conditional parallel trends assumption that

Coarse Time-Varying Conditional Parallel Trends:

$$E[Y_k(\infty) - Y_{k-1}(\infty)|T = m, A_m, \bar{L}_m] = E[Y_k(\infty) - Y_{k-1}(\infty)|T > m, \bar{L}_m] \forall k > m. \tag{10}$$

An implication of (10) is that

$$E[Y_k(\infty) - Y_{k-1}(\infty)|T \geq m, A_m = a_m, \bar{L}_m = \bar{l}_m] \tag{11}$$

does not depend on a_m. In words, conditional on observed covariate history through time m, the expected untreated counterfactual outcome trends from time m onwards are the same in those who initiate treatment (at any value) at time m and those who do not yet initiate treatment at time m. (10) will often be a more plausible assumption than a version that does not condition on time-varying covariates, e.g. Callaway and Sant'anna (2021). Note again that as in Remark 1 our definition of $\bar{L}_m$ implies that we do not condition on Y_m in (8) or (10).

3 Identification and g-Estimation of Coarse SNMMs under Parallel Trends

Because much of the DiD literature deals with the staggered adoption setting, where the positivity assumption is only satisfied for coarse SNMMs, we will discuss identification and estimation of coarse SNMMs first.

3.1 Nonparametric Identification

For any γ^c a vector of functions $\gamma^c_{mk}(\bar{l}_m, a_m)$ ranging over $K \geq k > m$, let

$$H^c_{mk}(\gamma^c) \equiv Y_k - \sum_{j=m}^{k-1} \gamma^c_{jk}(\bar{L}_j, A_j)\mathbb{1}\{T = j\} \text{ for } k > m \text{ and } H^c_{mm}(\gamma^c) \equiv Y_m. \quad (12)$$

$H^c_{mk}(\gamma^{c*})$ has the following important properties (Robins, 1998) for all $k > m$:

$$\begin{aligned} &E[H^c_{mk}(\gamma^{c*})|\bar{L}_m, T \geq m, A_m] = E[Y_k(\infty)|\bar{L}_m, T \geq m, A_m] \\ &E[H^c_{0k}(\gamma^{c*})] = E[Y_k(\infty)]. \end{aligned} \quad (13)$$

From (13) and the coarse time-varying conditional parallel trends assumption (10) it follows that

$$\begin{aligned} &E[H^c_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})|T \geq m, A_m = a_m, \bar{L}_m] = \\ &E[H^c_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})|T > m, \bar{L}_m] \ \forall k > m. \end{aligned} \quad (14)$$

Note that the quantity $H^c_{mk}(\gamma^{c*})$ behaves like the counterfactual quantity $Y_k(\infty)$ in that its conditional trend in those not treated prior to time m does not depend on A_m. We can exploit this property to identify and construct doubly robust estimating equations for γ^{c*} and other quantities of interest.

Theorem 1. *Under (1), (7), and (10),*
(i) γ^{c*} is identified from the joint distribution of $O = (\bar{L}_K, \overline{A}_K)$ as the unique solution to

$$E\left[\sum_{K \geq k > m} U^c_{mk}(s_m, \gamma^c)\right] = 0 \quad (15)$$

where

$$\begin{aligned} U^c_{mk}(s_m, \gamma^c) = \mathbb{1}\{T \geq m\}\{&H^c_{mk}(\gamma^c) - H^c_{m(k-1)}(\gamma^c)\}\times \\ &\{s_m(k, \overline{L}_m, A_m) - E\left[s_m(k, \overline{L}_m, A_m)|\overline{L}_m, T \geq m\right]\} \end{aligned}$$

as $s_m(k, \bar{l}_m, a_m)$ varies over all functions of $k, \bar{l}_m, a_m$. (If A_m are Bernoulli then the last expression is $s_m(k, \overline{L}_m, a_m = 1)\{A_m - E[A_m|\overline{L}_m, T \geq m]\}$.)

(ii) γ^{c*} satisfies the estimating function

$$E[\sum_{K \geq k > m} \sum_{m=0}^{K-1} U_{mk}^{c\dagger}(s_m, \gamma^{c*}, \widetilde{v}_m^c, \widetilde{E}_{A_m|\overline{L}_m, T\geq m})] = 0 \tag{16}$$

with

$$U_{mk}^{c\dagger}(s_m, \gamma^{c*}, \widetilde{v}_m^c, \widetilde{E}_{A_m|\bar{l}_m, T\geq m}) = \mathbb{1}\{T \geq m\}\{H_{mk}^c(\gamma^{c*}) - H_{m(k-1)}(\gamma^{c*}) - \widetilde{v}_m(k, \overline{L}_m)\} \times$$
$$\{s_m(k, \overline{L}_m, A_m) - \widetilde{E}_{A_m|\overline{L}_m, T\geq m}[s_m(k, \overline{L}_m, A_m)|\overline{L}_m, T \geq m]\}$$

for any $s_m(k, \bar{l}_m, a_m)$ if either
(a) $\widetilde{v}_m^c(k, \bar{l}_m) = v_m^c(k, \bar{l}_m; \gamma^{c*})$ for all $k > m$, where

$$v_m^c(k, \bar{l}_m; \gamma^c) \equiv E[H_{mk}^c(\gamma^c) - H_{mk-1}^c(\gamma^c)|\bar{L}_m, T \geq m]$$

or
(b) $\widetilde{E}_{A_m|\overline{L}_m, T\geq m} = E_{A_m|\overline{L}_m, T\geq m}$ for all m.

(16) is thus a doubly robust estimating function.

Proof. See Appendix A. □

It is of course an immediate corollary of Theorem 1 that quantities derived from γ^{c*}, such as $E[\bar{Y}_K(\infty)]$ and $E[\underline{Y}_{m+1}(\infty)|T = m, A_m = a_m]$ discussed in Section 2.2, are also identified.

3.2 Estimation

Theorem 1 establishes nonparametric identification of time-varying heterogeneous effects modeled by a coarse SNMM under a parallel trends assumption conditional on time-varying covariates. Estimation would often proceed by first specifying a model (9) with finite dimensional parameter ψ_c^*. It follows from the doubly robust estimating equations in Theorem 1 and arguments in Chernozhukov et al. (2018) and Smucler et al. (2019) that under regularity conditions we can obtain a consistent asymptotically normal estimator of ψ_c^* via a cross-fitting procedure. To do so, we shall need to estimate the unknown conditional means $v_m^c(k, \bar{l}_m; \psi_c)$ and $E_{A_m|\overline{L}_m, T\geq m}$ (hereafter nuisance functions) that are present in the $U_{mk}^{c\dagger}(s_m, \gamma^c(\psi_c), v_m^c(k, \bar{l}_m; \psi_c), E_{A_m|\overline{L}_m, T\geq m})$, where we now index $\gamma^c(\psi_c)$ and

$v_m^c(k,\bar{l}_m;\gamma^c(\psi_c))$ by the parameter ψ_c of our coarse SNMM. We will consider state-of-the-art cross-fit doubly robust machine learning (DR-ML) estimators (Chernozhukov et al., 2018; Smucler et al., 2019) of ψ_c^* in which the nuisance functions are estimated by arbitrary machine learning algorithms chosen by the analyst.

The following algorithm computes our cross fit estimator $\hat{\psi}_c^{cf} = \hat{\psi}_c^{cf}(s, \widehat{v}^c, \widehat{E}_{A_m|\bar{L}_m,T\geq m})$, where $s = (s_1, \ldots, s_K)$ is a user chosen vector of vector functions with range the dimension of ψ_c:

(i) Randomly split the N study subjects into 2 parts: an estimation sample of size n and a training (nuisance) sample of size $n_{tr} = N - n$ with $n/N \approx 1/2$. Without loss of generality we shall assume that $i = 1, \ldots, n$ corresponds to the estimation sample.

(ii) Applying machine learning methods to the training sample data, construct estimators $\hat{E}_{A_m|\bar{L}_m,T\geq m}$ of $E_{A_m|\bar{L}_m,T\geq m}$ and $\hat{v}_m^c(k,\bar{l}_m;\gamma^c(\psi_c))$ of $E[H_{mk}^c(\gamma^c(\psi_c)) - H_{mk-1}^c(\gamma^c(\psi_c))|\bar{L}_m, T \geq m]$.

(iii) Let $\hat{U}^{c\dagger}(s,\gamma^c(\psi_c)) = \sum_{K\geq k>m}\sum_{m=0}^{K-1}\hat{U}_{mk}^{c\dagger}(s_m,\gamma^c(\psi_c))$ where $\hat{U}_{mk}^{c\dagger}(s_m,\gamma^c(\psi_c))$ is defined to be $U_{mk}^{c\dagger}(s_m,\gamma^c(\psi_c))$ from (16) except with $\hat{E}_{A_m|\bar{L}_m,T\geq m}$ and $\hat{v}_m^c(k,\bar{l}_m;\gamma^c(\psi_c))$ obtained in (ii) substituted for $\widetilde{E}_{A_m|\bar{L}_m,T\geq m}$ and $\widetilde{v}_m^c(k,\bar{l}_m)$. Compute $\hat{\psi}_c^{(1)}$ from the n subjects in the estimation sample as the (assumed unique) solution to vector estimating equations $\mathbb{P}_n[\hat{U}^{c\dagger}(s,\gamma^c(\psi_c))] = 0$. Next, compute $\hat{\psi}_c^{(2)}$ just as $\hat{\psi}_c^{(1)}$, but with the training and estimation samples reversed.

(iv) Finally, the cross fit estimate $\hat{\psi}_c^{cf}$ is $(\hat{\psi}_c^{(1)} + \hat{\psi}_c^{(2)})/2$.

Remark 2. *Because step (ii) of the cross-fitting procedure estimates conditional expectations of functions of ψ_c, the estimator $\hat{\psi}_c^{(1)}$ must, in general, be solved iteratively and might be difficult to compute. Remark 20 in Appendix 5 of Liu et al (2021) discusses strategies for reducing the computational burden. If parameters $\psi_{c;mk}$ and $\psi_{c;m'k}$ of blip functions γ_{mk}^c and $\gamma_{m'k}^c$ are variation independent for $m \neq m'$, then it is possible to backwards recursively estimate $\psi_{c;mk}^*$ given estimates $\hat{\psi}_{c;m'k}$ for $m' > m$. If, for each j and k, $\gamma_{jk}^c(\bar{l}_j, a_j; \psi_{c;jk})$ is further assumed to be linear in $\psi_{c;jk}$, i.e. $\gamma_{jk}^c(\bar{L}_j, A_j; \psi_{c;jk}) = \psi_{c;jk}^T R_{jk}$ for a given vector transformation $R_{jk}(\bar{L}_j, A_j)$, then it is possible to estimate $E[H_{mk}^c(\gamma^c(\psi_c)) - H_{mk-1}^c(\gamma^c(\psi_c))|\bar{L}_m, T \geq m]$ by*

$$
\begin{aligned}
&E[H_{mk}^c(\gamma^c(\psi_{c;mk};\underline{\hat{\psi}}_{c;m+1,k})) - H_{mk-1}^c(\gamma^c(\psi_{c;mk-1};\underline{\hat{\psi}}_{c;m+1,k-1}))|\bar{L}_m, T \geq m] \\
&= E[(Y_k - \sum_{j=m+1}^{k}\gamma_{jk}^c(\bar{L}_j, A_j; \hat{\psi}_{c;jk})) - (Y_{k-1} - \sum_{j=m+1}^{k-1}\gamma_{jk-1}^c(\bar{L}_j, A_j; \hat{\psi}_{c;jk-1}))|\bar{L}_m, T \geq m] \\
&- \psi_{c;mk}^T E[R_{mk}|\bar{L}_m, T \geq m] + \psi_{c;mk-1}^T E[R_{mk-1}|\bar{L}_m, T \geq m]
\end{aligned}
$$

where $\underline{\hat{\psi}}_{c;m+1,k}$ denotes $(\hat{\psi}_{c;m+1,k}, \ldots, \hat{\psi}_{c;k,k})$. The resulting estimator never requires estimation of a conditional expectation of a function of a yet to be estimated $\psi_{c;mk}$ and is thus straightforward to compute. See Lewis and Syrgkanis (2020) for a related approach to estimation enabling lasso estimation of sparse high dimensional blip function parameters. If there is insufficient data to estimate variation independent blip function parameters at each time point, analysts might consider a parametric approach, which we discuss in Remarks 4 and 5 below.

Theorem 2. *The nuisance training sample is denoted as Nu. If (a) $E[\hat{U}^{c\dagger}(s, \gamma^c(\psi_c^*))|Nu]$ is $o_p(n^{-1/2})$ and (b) all of the estimated nuisance conditional expectations and densities converge to their true values in $L_2(P)$, then*

$$\hat{\psi}_c^{(1)} - \psi_c^* = n^{-1} \sum_{i=1}^{n} IF_i^c + o_p(n^{-1/2})$$

$$\hat{\psi}_c^{cf} - \psi_c^* = n^{-1} \sum_{i=1}^{n} IF_i^c + o_p(n^{-1/2})$$

where $IF^c \equiv IF^c(s, \psi_c^*) = J_c^{-1} U_{mk}^{c\dagger}(s, \gamma^c(\psi_c^*), v_m^c(\gamma^c(\psi_c^*)), E_{A_m|\bar{L}_m, T \geq m}) - \psi_c^*$ is the influence function of $\hat{\psi}_c^{cf}$ and $\hat{\psi}_c^{(1)}$ and

$$J_c = \frac{\partial}{\partial \psi_c^T} E[U_{mk}^{c\dagger}(s, \gamma^c(\psi_c), v_m^c(\gamma^c(\psi_c)), E_{A_m|\bar{L}_m, T \geq m})]|_{\psi_c = \psi_c^*}.$$

Further, $\hat{\psi}_c^{cf}$ is a regular asymptotically linear estimator of ψ_c^* with asymptotic variance equal to $var[IF^c]$.

Proof. The results follow from Theorem 21 in Appendix 5 of Liu et al (2021). □

Remark 3. *A sufficient condition (Smucler et al., 2019) for $E[\hat{U}^{c\dagger}(s, \gamma^c(\psi_c^*))|Nu]$ to be $o_p(n^{-1/2})$ is that*

$$max_{k>m}\{||\hat{v}_m^c(k, \bar{l}_m; \gamma^c(\psi_c^*)) - v_m^c(k, \bar{l}_m; \gamma^c(\psi_c^*))|| \times$$
$$||\hat{E}_{A_m|\overline{L}_m, T \geq m}[s_m(k, \overline{L}_m, A_m)|\overline{L}_m, T \geq m] - E_{A_m|\overline{L}_m, T \geq m}[s_m(k, \overline{L}_m, A_m)|\overline{L}_m, T \geq m]||\}$$
$$= o_p(n^{-1/2}).$$

That is, for every $k > m$ the product of the rates of convergence of the nuisance estimators for $v_m^c(k, \bar{l}_m; \gamma^c(\psi_c^*))$ and $E_{A_m|\overline{L}_m, T \geq m}$ is $o_p(n^{-1/2})$. This property is referred to as rate double robustness by Smucler et al. (2019).

It follows from (13) and Theorems 1 and 2, that we can also construct consistent and asymptotically normal (CAN) plug-in estimators of other quantities of interest using $\gamma^c(\hat{\psi}_c^{cf})$. $E_{\hat{\psi}_c^{cf}}[Y_k(\infty)] = \mathbb{P}[H_{0k}^c(\gamma^c(\hat{\psi}_c^{cf}))]$ is a CAN estimator for $E[Y_k(\infty)]$. $E_{\hat{\psi}_c^{cf}}[Y_k(\infty)|T = m] = \mathbb{P}^{T=m}[H_{mk}^c(\gamma_{mk}^c(\hat{\psi}_c^{cf}))]$ is a CAN estimator for $E[Y_k(\infty)|T = m]$ where $\mathbb{P}^{T=m}$ denotes sample average among subjects with $T = m$. $E_{\hat{\psi}_c^{cf}}[Y_k(\infty)|T = m, \bar{L}_m \in B] = \mathbb{P}^{T=m,B}[H_{mk}^c(\gamma_{mk}^c(\hat{\psi}_c^{cf}))]$ is a CAN estimator for $E[Y_k(\infty)|T = m, \bar{L}_m \in B]$ where $\mathbb{P}^{T=m,B}$ denotes sample average among subjects with $T = m$ and $\bar{L}_m \in B$ where B is an event in the sample space of $\bar{L}_m$ with positive probability. For example, B could be the event that the j^{th} covariate L_m^j in the vector L_m measured at time m exceeds a certain threshold h (i.e. $\bar{L}_m \in B$ if $L_m^j > h$). Confidence intervals for all of these quantities can be computed via nonparametric bootstrap. (In Appendix B, we revisit the hypothetical analysis adjusting for non-adherence in a randomized trial with repeated outcome measurements. This hypothetical example illustrates how the quantities described above map onto substantive questions that can be addressed via a coarse SNMM. The example also illustrates some advantages of fitting a coarse SNMM assuming parallel trends conditional on time-varying covariates versus a traditional DiD approach assuming parallel trends conditional on baseline covariates alone.)

Remark 4. *Alternatively, one can specify parametric nuisance models $E_{A_m|\bar{L}_m,T\geq m;\rho}$ and $v_m(k, \bar{l}_m; \gamma^c(\psi_c); \phi(\psi_c))$ for $E_{A_m|\bar{L}_m,T\geq m}$ and $E[H_{mk}^c(\gamma^c(\psi_c)) - H_{mk-1}^c(\gamma^c(\psi_c))|\bar{L}_m, T \geq m]$, respectively, and forego cross fitting. Under the assumptions of Theorem 1 and standard regularity conditions for theory of M-estimators (e.g. van der Vaart, 1998, Chapter 5), if at least one parametric nuisance model along with (9) is correctly specified and consistently estimated at $n^{-1/2}$ rate, then the estimator $\tilde{\psi}_c$ solving estimating equations $\mathbb{P}_n[\hat{U}^{c\dagger}(s, \gamma^c(\psi_c))] = 0$ with $E_{A_m|\bar{L}_m,\bar{A}_{m-1};\hat{\rho}}$ and $v_m(k, \bar{l}_m; \gamma^c(\psi_c); \hat{\phi}(\psi_c))$ in place of $\hat{E}_{A_m|\bar{L}_m,\bar{A}_{m-1}}$ and $\hat{v}_m^c(k, \bar{l}_m; \gamma^c(\psi_c); \hat{\phi}(\psi_c))$, respectively, is consistent and asymptotically normal and confidence intervals for ψ_c^* and many derived quantities of interest can be computed via bootstrap.*

Remark 5. *Suppose the blip model is specified to be linear in ψ_c (i.e. $\gamma_{mk}^c(a_m, \bar{l}_m) = a_m\psi_c^T R_{mk}(\bar{l}_m)$ for $R_{mk}(\bar{l}_m)$ some transformation of history through time m the dimension of ψ_c) and the nuisance model $v_m^c(k, \bar{l}_m, \gamma^c(\psi_c); \phi(\psi_c))$ is specified to be linear in ϕ (i.e. $E[H_{mk}^c(\gamma(\psi)) - H_{mk-1}^c(\gamma^c(\psi_c))|\bar{L}_m, T \geq m] = \phi^T D_{mk}(\bar{l}_m)$ for $D_{mk}(\bar{l}_m)$ some transformation of history through time m the dimension of ϕ). Then the doubly robust estimator $(\hat{\psi}_c, \hat{\phi})$ is available in closed form as*

$$
\begin{aligned}
(\hat{\psi}_c, \hat{\phi})^T = (&\sum_i \sum_{k>m} \mathbb{1}\{T \geq m\}(Y_{ik} - Y_{ik-1}) \begin{pmatrix} s_{im}X_{im} \\ D_{im} \end{pmatrix}) \times \\
(&\sum_i \sum_{k>m} \mathbb{1}\{T \geq m\}(V_{imk} - V_{imk-1}, D_{im}) \begin{pmatrix} s_{im}X_{im} \\ D_{im} \end{pmatrix})^{-1},
\end{aligned} \tag{17}
$$

where $V_{imk} = \sum_{j=m}^{k} A_{ij} R_{jk}$*,* $X_{im} = A_{im} - \hat{E}[A_m|\bar{L}_{im}, T_i \geq m]$*, and* s_{im} *is the usual analyst specified vector with dimension equal to the dimension of* ψ_c*.*

4 Identification and g-Estimation of Standard SNMMs Under Parallel Trends

Standard SNMMs can model certain effect contrasts that coarse SNMMs cannot, and in circumstances when they are both reasonable options Robins (1998) argues that parametric standard SNMMs can better encode prior knowledge than parametric coarse SNMMs. Thus, it is also important to consider identification and estimation of standard SNMMs.

4.1 Nonparametric Identification

Let

$$H_{mk}(\gamma) \equiv Y_k - \sum_{j=m}^{k-1} \gamma_{jk}(\bar{A}_j, \bar{L}_j) \text{ for } k > m \text{ and } H_{tt} \equiv Y_t. \tag{18}$$

Robins (1994) showed that $H_{mk}(\gamma^*)$ has the following important properties for all $k > m$:

$$\begin{aligned} &E[H_{mk}(\gamma^*)|\bar{L}_m, \bar{A}_m] = E[Y_k(\bar{A}_{m-1}, \underline{0})|\bar{L}_m, \bar{A}_m] \\ &E[H_{0k}(\gamma^*)] = E[Y_k(\bar{0})]. \end{aligned} \tag{19}$$

By (19) and the time-varying conditional parallel trends assumption (5), it follows that

$$E[H_{mk}(\gamma^*) - H_{mk-1}(\gamma^*)|\bar{A}_m = \bar{a}_m, \bar{L}_m] = E[H_{mk}(\gamma^*) - H_{mk-1}(\gamma^*)|\bar{A}_m = (\bar{a}_{m-1}, 0), \bar{L}_m] \forall k > m. \tag{20}$$

That is, given the true SNMM blip function, the observable quantity $H_{mk}(\gamma^*)$ behaves like the counterfactual quantity $Y_k(\bar{A}_{m-1}, \underline{0})$ in that its conditional trend does not depend on A_m. We can again exploit this property to identify and construct doubly robust estimating equations for γ^*.

Theorem 3. *Under (1), (2), and (5),*
(i) γ^* *is identified from the joint distribution of* $O = (\bar{L}_K, \overline{A}_K)$ *as the unique solution to*

$$E\left[\sum_{K \geq k > m} U_{mk}(s_m, \gamma)\right] = 0 \tag{21}$$

where

$$\begin{aligned} U_{mk}(s_m, \gamma) = \{&H_{mk}(\gamma) - H_{m(k-1)}(\gamma)\}\times \\ &\{s_m(k, \overline{L}_m, \bar{A}_m) - E[s_m(k, \overline{L}_m, \bar{A}_m)|\overline{L}_m, \bar{A}_{m-1}]\} \end{aligned}$$

as $s_m(k, \bar{l}_m, \bar{a}_m)$ varies over all functions of $k, \bar{l}_m, \bar{a}_m$. (If A_m are Bernoulli then the last expression is $s_m(k, \overline{L}_m, \overline{A}_{m-1}, a_m = 1)\{A_m - E[A_m|\overline{L}_m, \overline{A}_{m-1}]\}$.)

(ii) $\gamma*$ satisfies the estimating function

$$E[\sum_{K \geq k > m} \sum_{m=0}^{K-1} U^{\dagger}_{mk}(s_m, \gamma^*, \widetilde{v}_m, \widetilde{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}})] = 0 \tag{22}$$

with

$$U^{\dagger}_{mk}(s_m, \gamma^*) = \{H_{mk}(\gamma^*) - H_{m(k-1)}(\gamma^*) - \widetilde{v}_m(k, \overline{L}_m, \overline{A}_{m-1})\}\times$$
$$\{s_m(k, \overline{L}_m, \overline{A}_m) - \widetilde{E}_{A_m|\overline{L}_m, \overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\overline{L}_m, \overline{A}_{m-1}]\}$$

for any $s_m(k, \bar{l}_m, \bar{a}_m)$ if either
(a) $\widetilde{v}_m(k, \bar{l}_m, \bar{a}_{m-1}) = v_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma^*)$ for all $k > m$, where

$$v_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma) \equiv E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{L}_m, \bar{A}_{m-1}]$$

or
(b) $\widetilde{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}} = E_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ for all m.

(22) is thus a doubly robust estimating function.

4.2 Estimation

Theorem 3 establishes nonparametric identification of time-varying heterogeneous effects modeled by a standard SNMM under a parallel trends assumption conditional on time-varying covariates. Estimation would often proceed by first specifying a model (4) with finite dimensional parameter ψ^*. It follows from the doubly robust estimating equations in Theorem 3 and arguments in Chernozhukov et al. (2018) and Smucler et al. (2019) that under regularity conditions we can obtain a consistent asymptotically normal estimator of ψ^* via a cross-fitting procedure. To do so, we shall need to estimate the unknown conditional means $v_m(k, \bar{l}_m, \bar{a}_{m-1}; \psi)$ and $E_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ (hereafter nuisance functions) that are present in the $U^{\dagger}_{mk}(s_m, \gamma(\psi), v_m(k, \bar{l}_m, \bar{a}_{m-1}; \psi), E_{A_m|\overline{L}_m, \bar{A}_{m-1}})$, where we now index $\gamma(\psi)$ and $v_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi))$ by the parameter ψ of our standard SNMM. We will consider state-of-the-art cross-fit doubly robust machine learning (DR-ML) estimators (Chernozhukov et al., 2018; Smucler et al., 2019) of ψ^* in which the nuisance functions are estimated by arbitrary machine learning algorithms chosen by the analyst.

The following algorithm computes our cross fit estimator $\hat{\psi}^{cf} = \hat{\psi}^{cf}(s, \widehat{v}, \widehat{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}})$, where $s = (s_1, \ldots, s_K)$ is a user chosen vector of vector functions with range the dimension of ψ:

(i) Randomly split the N study subjects into two parts: an estimation sample of size n and a training (nuisance) sample of size $n_{tr} = N - n$ with $n/N \approx 1/2$. Without loss of generality we shall assume that $i = 1, \ldots, n$ corresponds to the estimation sample.

(ii) Applying machine learning methods to the training sample data, construct estimators $\hat{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ of $E_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ and $\hat{v}_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi))$ of $E[H_{mk}(\gamma(\psi)) - H_{mk-1}(\gamma(\psi))|\bar{L}_m, \bar{A}_{m-1}]$.

(iii) Let $\hat{U}^{\dagger}(s, \gamma(\psi)) = \sum_{K \geq k > m} \sum_{m=0}^{K-1} \hat{U}^{\dagger}_{mk}(s_m, \gamma(\psi))$ where $\hat{U}^{\dagger}_{mk}(s_m, \gamma(\psi))$ is defined to be $U^{\dagger}_{mk}(s_m, \gamma(\psi))$ from (22) except with $\hat{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ and $\hat{v}_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi))$ obtained in (ii) substituted for $\widetilde{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ and $\widetilde{v}_m(k, \bar{l}_m, \bar{a}_{m-1})$. Compute $\hat{\psi}^{(1)}$ from the n subjects in the estimation sample as the (assumed unique) solution to vector estimating equations $\mathbb{P}_n[\hat{U}^{\dagger}(s, \gamma(\psi))] = 0$. Next, compute $\hat{\psi}^{(2)}$ just as $\hat{\psi}^{(1)}$, but with the training and estimation samples reversed.

(iv) Finally, the cross fit estimate $\hat{\psi}^{cf}$ is $(\hat{\psi}^{(1)} + \hat{\psi}^{(2)})/2$.

Remark 6. *Because step (ii) of the cross-fitting procedure estimates conditional expectations of functions of ψ, the estimator $\hat{\psi}^{(1)}$ must, in general, be solved iteratively and might be difficult to compute. Remark 2 above and Remark 20 in Appendix 5 of Liu et al (2021) discusses strategies for reducing the computational burden. If parameters ψ_{mk} and $\psi_{m'k}$ of blip functions γ_{mk} and $\gamma_{m'k}$ are variation independent for $m \neq m'$, then it is possible to backwards recursively estimate ψ^*_{mk} given estimates $\hat{\psi}_{m'k}$ for $m' > m$. If, for each j and k, $\gamma_{jk}(\bar{l}_j, \bar{a}_j; \psi_{jk})$ is further assumed to be linear in ψ_{jk}, i.e. $\gamma_{jk}(\bar{L}_j, \bar{A}_j; \psi_{jk}) = \psi^T_{jk} R_{jk}$ for a given vector transformation $R_{jk}(\bar{L}_j, \bar{A}_j)$, then it is possible to estimate $E[H_{mk}(\gamma(\psi)) - H_{mk-1}(\gamma(\psi))|\bar{L}_m, \bar{A}_{m-1}]$ by*

$$
\begin{aligned}
&E[H_{mk}(\gamma(\psi_{mk}; \underline{\hat{\psi}}_{m+1,k})) - H_{mk-1}(\gamma(\psi_{mk-1}; \underline{\hat{\psi}}_{m+1,k-1}))|\bar{L}_m, \bar{A}_{m-1}] \\
&= E[(Y_k - \sum_{j=m+1}^{k} \gamma_{jk}(\bar{L}_j, \bar{A}_j; \hat{\psi}_{jk})) - (Y_{k-1} - \sum_{j=m+1}^{k-1} \gamma_{jk-1}(\bar{L}_j, \bar{A}_j; \hat{\psi}_{jk-1}))|\bar{L}_m, \bar{A}_{m-1}] \\
&- \psi^T_{mk} E[R_{mk}|\bar{L}_m, \bar{A}_{m-1}] + \psi^T_{mk-1} E[R_{mk-1}|\bar{L}_m, \bar{A}_{m-1}]
\end{aligned}
$$

where $\underline{\hat{\psi}}_{m+1,k}$ denotes $(\hat{\psi}_{m+1,k}, \ldots, \hat{\psi}_{k,k})$. The resulting estimator never requires estimation of a conditional expectation of a function of a yet to be estimated ψ_{mk} and is thus straightforward to compute. See Lewis and Syrgkanis (2020) for a related *approach to estimation enabling lasso estimation of sparse high dimensional blip function parameters. If there is insufficient data to estimate variation independent blip function parameters at each time point, analysts might consider a parametric approach, which we discuss in Remarks 8 and 9 below.*

Theorem 4. *The nuisance training sample is denoted as Nu. If (a) $E[\hat{U}^{\dagger}(s, \gamma(\psi^{*}))|Nu]$ is $o_p(n^{-1/2})$ and (b) all of the estimated nuisance conditional expectations and densities converge to their true values in $L_2(P)$, then*

$$\hat{\psi}^{(1)} - \psi^{*} = n^{-1}\sum_{i=1}^{n} IF_i + o_p(n^{-1/2})$$

$$\hat{\psi}^{cf} - \psi^{*} = n^{-1}\sum_{i=1}^{n} IF_i + o_p(n^{-1/2})$$

where $IF \equiv IF(s, \psi^{*}) = J^{-1}U^{\dagger}_{mk}(s, \gamma(\psi^{*}), v_m(\gamma(\psi^{*})), E_{A_m|\bar{L}_m, \bar{A}_{m-1}}) - \psi^{*}$ is the influence function of $\hat{\psi}^{cf}$ and $\hat{\psi}^{(1)}$ and $J = \frac{\partial}{\partial \psi^{T}} E[U^{\dagger}_{mk}(s, \gamma(\psi), v_m(\gamma(\psi)), E_{A_m|\bar{L}_m, \bar{A}_{m-1}})]|_{\psi=\psi^{*}}$. Further, $\hat{\psi}^{cf}$ is a regular asymptotically linear estimator of ψ^{*} with asymptotic variance equal to $var[IF]$.

Proof. The results follow from Theorem 5 of Liu et al (2021). □

Remark 7. *A sufficient condition (Smucler et al., 2019) for $E[\hat{U}^{\dagger}(s, \gamma(\psi^{*}))|Nu]$ to be $o_p(n^{-1/2})$ is that*

$$max_{k>m}\{||\hat{v}_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi^{*})) - v_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi^{*}))|| \times$$
$$||\hat{E}_{A_m|\overline{L}_m, \bar{A}_{m-1}}[s_m(k, \overline{L}_m, \bar{A}_m)|\overline{L}_m, \bar{A}_{m-1}] - E_{A_m|\overline{L}_m, \bar{A}_{m-1}}[s_m(k, \overline{L}_m, \bar{A}_m)|\overline{L}_m, \bar{A}_{m-1}]||\}$$
$$= o_p(n^{-1/2}).$$

That is, for every $k > m$ the product of the rates of convergence of the nuisance estimators for $v_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi^{*}))$ and $E_{A_m|\overline{L}_m, \bar{A}_{m-1}}$ is $o_p(n^{-1/2})$. This property is referred to as rate double robustness by Smucler et al. (2019).

It follows from (19) and Theorems 3 and 4, that we can also construct CAN plug-in estimators of other quantities of interest using $\gamma(\hat{\psi}^{cf})$. $E_{\hat{\psi}^{cf}}[Y_k(\bar{0})] = \mathbb{P}[H_{0k}(\gamma(\hat{\psi}^{cf}))]$ is a CAN estimator for $E[Y_k(\infty)]$ where $\mathbb{P}$ denotes sample average. $E_{\hat{\psi}^{cf}}[Y_k(\bar{A}_{m-1}, \underline{0})] = \mathbb{P}[H_{mk}(\gamma_{mk}(\hat{\psi}^{cf}))]$ is a CAN estimator for $E[Y_k(\bar{A}_{m-1}, \underline{0})]$. $E_{\hat{\psi}^{cf}_c}[Y_k(\bar{A}_{m-1}, \underline{0})|(\bar{A}_{m-1}, \bar{L}_m) \in B] = \mathbb{P}^{B}[H_{mk}(\gamma_{mk}(\hat{\psi}^{cf}))]$ is a CAN estimator for $E[Y_k(\bar{A}_{m-1}, \underline{0})|(\bar{A}_{m-1}, \bar{L}_m) \in B]$ where $\mathbb{P}^{B}$ denotes sample average among subjects with $(\bar{A}_{m-1}, \bar{L}_m) \in B$ where B is an event in the sample space of treatment and covariate history $(\bar{A}_{m-1}, \bar{L}_m)$ with positive probability. For example, B could be the event that there was no treatment through $m-1$ and the average value of another covariate through time m exceeds some threshold. Of course, if we want to know the effect at a particular history $\bar{L}_m = \bar{l}_m$, we can just query the blip function directly to consistently estimate $E[Y_k(\bar{A}_{m-1} = 0, 1, \underline{0})|\bar{A}_{m-1} = 0, A_m = 1, \bar{L}_m = \bar{l}_m] - E[Y_k(\bar{0})|\bar{A}_{m-1} = 0, A_m = 1, \bar{L}_m = \bar{l}_m]$ by $\gamma(\bar{a}_{m-1} = 0, a_m = 1, \bar{l}_m; \hat{\psi}^{cf})$. Previously existing DiD methods cannot estimate the blip function of a standard SNMM or certain derived quantities of interest, but our estimators can. There are many other effect contrasts that might be of interest in specific applications that could be identified under standard SNMMs (and therefore under parallel trends assumptions using our methods) but not under coarse SNMMs or using previously existing DiD methods. Confidence intervals for all of these quantities can be computed via nonparametric bootstrap.

Remark 8. *Alternatively, one can specify parametric nuisance models* $E_{A_m|\bar{L}_m,\bar{A}_{m-1};\rho}$ *and* $v_m(k, \bar{l}_m, \bar{a}_{m-1}, \gamma(\psi); \phi(\psi))$ *for* $E_{A_m|\bar{L}_m,\bar{A}_{m-1}}$ *and* $E[H_{mk}(\gamma(\psi)) - H_{mk-1}(\gamma(\psi))|\bar{L}_m, \bar{A}_{m-1}]$, *respectively, and forego cross fitting. Under the assumptions of Theorem 3 and standard regularity conditions for theory of M-estimators (e.g. van der Vaart, 1998, Chapter 5), if at least one parametric nuisance model along with (4) is correctly specified and consistently estimated at* $n^{-1/2}$ *rate, then the estimator* $\tilde{\psi}$ *solving estimating equations* $\mathbb{P}_n[\hat{U}^\dagger(s, \gamma(\psi))] = 0$ *with* $\hat{E}_{A_m|\bar{L}_m,\bar{A}_{m-1}} = E_{A_m|\bar{L}_m,\bar{A}_{m-1};\hat{\rho}}$ *and* $\hat{v}_m = v_m(k, \bar{l}_m, \bar{a}_{m-1}; \gamma(\psi); \hat{\phi}(\psi))$ *is consistent and asymptotically normal and confidence intervals for* ψ^* *and many derived quantities of interest can be computed via bootstrap.*

Remark 9. *Suppose the blip model is specified to be linear in* ψ *(i.e.* $\gamma_{mk}(\bar{a}_m, \bar{l}_m) = a_m \psi^T R_{mk}(\bar{a}_{m-1}, \bar{l}_m)$ *for* $R_{mk}(\bar{a}_{m-1}, \bar{l}_m)$ *some transformation of history through time* m *the dimension of* ψ*) and the nuisance model* $v_m(k, \bar{l}_m, \bar{a}_{m-1}, \gamma(\psi); \phi(\psi))$ *is specified to be linear in* ϕ *(i.e.* $E[H_{mk}(\gamma(\psi)) - H_{mk-1}(\gamma(\psi))|\bar{L}_m, \bar{A}_{m-1}] = \phi^T D_{mk}(\bar{a}_{m-1}, \bar{l}_m)$ *for* $D_{mk}(\bar{a}_{m-1}, \bar{l}_m)$ *some transformation of history through time* m *the dimension of* ϕ*). Then the doubly robust estimator* $(\hat{\psi}, \hat{\phi})$ *is available in closed form as*

$$(\hat{\psi}, \hat{\phi})^T = (\sum_i \sum_{k>m} (Y_{ik} - Y_{ik-1}) \begin{pmatrix} s_{im} X_{im} \\ D_{im} \end{pmatrix})(\sum_i \sum_{k>m} (V_{imk} - V_{imk-1}, D_{im}) \begin{pmatrix} s_{im} X_{im} \\ D_{im} \end{pmatrix})^{-1}, \tag{23}$$

where $V_{imk} = \sum_{j=m}^{k} A_{ij} R_{jk}$, $X_{im} = A_{im} - \hat{E}[A_m|\bar{L}_{im}, \bar{A}_{im-1}]$, *and* s_{im} *is the usual analyst specified vector with dimension equal to the dimension of* ψ.

5 Multiplicative Effects

The SNMM framework also readily handles multiplicative effects when the parallel trends assumption is assumed to hold on the additive scale, a scenario that has caused some consternation in the DiD literature (Ciani and Fisher, 2018). Suppose we are in the general treatment pattern setting. Define the multiplicative causal contrasts

$$e^{\gamma^{\times *}_{mk}(\bar{l}_m, \bar{a}_m)} \equiv \frac{E[Y_k(\bar{a}_m, \underline{0})|\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m]}{E[Y_k(\bar{a}_{m-1}, \underline{0})|\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m]} \tag{24}$$

for $k > m$. $\gamma^{\times *}_{mk}(\bar{a}_m, \bar{l}_m)$ is the average multiplicative effect at time k among patients with history $(\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m)$ of receiving treatment a_m at time m and then 0 thereafter compared to receiving treatment 0 at time m and thereafter. (Here '0' can be replaced by any baseline value a^*_m.)

A parametric multiplicative SNMM imposes functional forms on the multiplicative blip functions $\gamma^{\times *}_{mk}(\bar{A}_m, \bar{L}_m)$ for each $k > m$, i.e.

$$e^{\gamma^{\times *}_{mk}(\bar{a}_m, \bar{l}_m)} = e^{\gamma^{\times}_{mk}(\bar{a}_m, \bar{l}_m; \psi^*_{\times})} \tag{25}$$

where $\psi^*_{\times}$ is an unknown parameter vector and $\gamma^{\times}_{mk}(\bar{a}_m, \bar{l}_m; \psi_{\times})$ is a known function equal to 0 whenever $\psi = 0$ or $a_m = 0$.

We make the same parallel trends assumption (5) as in the additive standard setting. (We can of course also analogously estimate multiplicative effects in the coarse setting.)

Let

$$H^{\times}_{mk}(\psi) \equiv Y_k exp\{-\sum_{j=m}^{k-1} \gamma^{\times *}_{jk}(\bar{A}_j, \bar{L}_j)\} \text{ for } k > m \text{ and } H^{\times}_{tt} \equiv Y_t. \tag{26}$$

Robins (1994) showed that $H^{\times}_{mk}(\gamma^{\times *})$ has the following important properties for all $k > m$:

$$E[H^{\times}_{mk}(\gamma^{\times *})|\bar{L}_m, \bar{A}_m] = E[Y_k(\bar{A}_{m-1}, \underline{0})|\bar{L}_m, \bar{A}_m] \tag{27}$$

$$E[H^{\times}_{0k}(\gamma^{\times *})] = E[Y_k(\bar{0})]. \tag{28}$$

By (27) and the time-varying conditional parallel trends assumption (5), it follows that

$$\begin{aligned} &E[H^{\times}_{mk}(\gamma^{\times *}) - H^{\times}_{mk-1}(\gamma^{\times *})|\bar{A}_m = \bar{a}_m, \bar{L}_m] = \\ &E[H^{\times}_{mk}(\gamma^{\times *}) - H^{\times}_{mk-1}(\gamma^{\times *})|\bar{A}_m = (\bar{a}_{m-1}, 0), \bar{L}_m] \\ &\forall k > m. \end{aligned} \tag{29}$$

That is, given the true multiplicative SNMM blip function, the quantity $H^{\times}_{mk}(\gamma^{\times *})$ behaves like the counterfactual quantity $Y_k(\bar{A}_{m-1}, \underline{0})$ in that its conditional trend does not depend on A_m. We can yet again exploit this property to identify and construct doubly robust estimating equations for $\gamma^{\times *}$ and various derivative quantities of interest. Identification and estimation theorems and proofs are identical to Section (4) with $H^{\times}_{mk}$ in place of H_{mk}, $\gamma^{\times *}$ in place of γ^*, $\psi^*_{\times}$ in place of ψ^*, and (27) in place of (19). We can also analogously define, identify, and estimate multiplicative coarse SNMMs.

6 General Regime and Optimal Regime SNMMs

6.1 g-Estimation of General Regime SNMMs Under Parallel Trends

Let $Y_k(g)$ denote the counterfactual value of the outcome at time k under possibly dynamic treatment regime $g \in \mathcal{G}$, where $g \equiv (g_0, g_1, \ldots, g_K)$ is a vector of functions $g_t : (\bar{L}_t, \bar{A}_{t-1}) \to a_t$ that determine treatment values at each time point given observed history and $\mathcal{G}$ is the set of all such treatment regimes. Robins (2004) explained that there is nothing special about $\bar{0}$ as the regime that is followed after a blip of treatment in a SNMM. We can instead define the blip functions relative to arbitrary regime g as

$$\gamma^{g*}_{mk}(\bar{l}_m, \bar{a}_m) \equiv E[Y_k(\bar{a}_m, \underline{g}_{m+1}) - Y_k(\bar{a}_{m-1}, \underline{g}_m)|\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m]. \tag{30}$$

$\gamma^{g}_{mk}(\bar{a}_m, \bar{l}_m)$ is the average effect at time k among patients with history $(\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m)$ of receiving treatment a_m at time m and then following regime g thereafter compared to following regime g from time m onward.

A parametric general regime SNMM imposes functional forms on $\gamma^{g}_{mk}(\bar{A}_m, \bar{L}_m)$ for each $k > m$, i.e.

$$\gamma^{g*}_{mk}(\bar{a}_m, \bar{l}_m) = \gamma^{g}_{mk}(\bar{a}_m, \bar{l}_m; \psi^*_g) \tag{31}$$

where ψ^*_g is an unknown parameter vector and $\gamma^{g}_{mk}(\bar{a}_m, \bar{l}_m; \psi_g)$ is a known function equal to 0 whenever $\psi_g = 0$ or $a_m = 0$.

We can then assume the time-varying conditional parallel trends assumption holds for g, i.e.

$$\begin{aligned} &E[Y_k(\bar{a}_{m-1}, \underline{g}_m) - Y_{k-1}(\bar{a}_{m-1}, \underline{g}_m)|\bar{A}_m = \bar{a}_m, \bar{L}_m] = \\ &E[Y_k(\bar{a}_{m-1}, \underline{g}_m) - Y_{k-1}(\bar{a}_{m-1}, \underline{g}_m)|\bar{A}_m = (\bar{a}_{m-1}, g(\bar{L}_m, \bar{a}_{m-1})), \bar{L}_m] \\ &\forall k > m. \end{aligned} \tag{32}$$

Next, define

$$H_{mk}(\gamma^{g*}) \equiv Y_k - \sum_{j=m}^{k-1} \gamma^{g*}_{jk}(\bar{L}_j, \bar{A}_j). \tag{33}$$

As before, due to the fact that $H_{mk}(\gamma^{g*})$ behaves like a counterfactual in conditional expectation and satisfies (19) (Robins, 2004), it follows that under (32)

$$\begin{aligned}&E[H_{mk}(\gamma^{g*}) - H_{mk-1}(\gamma^{g*})|\bar{A}_m = \bar{a}_m, \bar{L}_m] = \\ &E[H_{mk}(\gamma^{g*}) - H_{mk-1}(\gamma^{g*})|\bar{A}_m = (\bar{a}_{m-1}, g(\bar{l}_m, \bar{a}_{m-1})), \bar{L}_m] \\ &\forall k > m.\end{aligned} \tag{34}$$

Thus, we can once again exploit this key property and follow exactly the same steps as in previous sections to identify and construct consistent doubly robust estimators for γ^{g*} and by extension $E[Y_k(g)]$.

The ability to estimate expected counterfactual outcomes under general dynamic treatment regimes under simple parallel trends assumptions might appear very exciting at first glance. But why should the parallel trends assumption hold for the particular regime g of interest? It seems difficult to justify this assumption for one regime over another. For this reason, in most practical settings, if one were to make this assumption for any given regime we would argue that one is effectively making the assumption for *all* regimes. Assuming parallel trends for all regimes, however, is a stronger assumption than is usually made in the DiD literature and effectively imposes restrictions on effect heterogeneity. In Appendix C, we show that parallel trends under all regimes (*practically*) implies that there is no effect modification by unobserved confounders, which we show in the following subsection enables identification of optimal regime SNMMs.

6.2 g-estimation of Optimal Regime SNMMs Under Parallel Trends and No Effect Modification By Unobserved Confounders

We assume that parallel trends (32) holds for all $g \in \mathcal{G}$. Let $\bar{U}_K$ denote a time-varying unobserved confounder, possibly multivariate and containing continuous and/or discrete components. Suppose the causal ordering at time m is (Y_m, U_m, L_m, A_m). We assume that were we able to observe $\bar{U}_K$ in addition to the observed variables we could adjust for all confounding, i.e.

$U-$**Sequential Exchangeability:** $A_m \perp\!\!\!\perp \underline{Y}_{m+1}(\bar{a}_K)|\bar{L}_m, \bar{A}_{m-1} = \bar{a}_{m-1}, \bar{U}_m \; \forall \; m \leq K, \bar{a}_K.$ (35)

We further assume

No Additive Effect Modification by U:

$$\begin{aligned}&\gamma^g_{mk}(\bar{l}_m, \bar{a}_m) = \gamma^g_{mk}(\bar{l}_m, \bar{a}_m, \bar{u}_m) \equiv \\ &E[Y_k(\bar{a}_m, \underline{g}_{m+1}) - Y_k(\bar{a}_{m-1}, \underline{g}_m)|\bar{A}_m = \bar{a}_m, \bar{L}_m = \bar{l}_m, \bar{U}_m = \bar{u}_m]\end{aligned} \tag{36}$$

for all regimes $g \in \mathcal{G}$. (It can actually be shown that if (36) holds for any single $g \in \mathcal{G}$, such as $\bar{0}$ then it holds for all $g \in \mathcal{G}$.) In Appendix C, we show that (36) practically, though not strictly mathematically, follows from (32) for all g.

Suppose we want to maximize the expectation of some utility $Y = \sum_{m=0}^{K} \tau_m Y_m$ which is a weighted sum of the outcomes we observe at each time step with weights τ_m. Let $Y(g) = \sum_{m=0}^{K} \tau_m Y_m(g)$ denote the counterfactual value of the utility under treatment regime g. We want to find $g^{opt} = argmax_g E[Y(g)]$. Under (35) and (36), it follows from results in Robins (2004) that g^{opt} is given by the following backward recursion:

$$g_K^{opt}(\bar{L}_K, \bar{A}_{K-1}) = argmax_{a_K} E[Y_{K+1}(\bar{A}_{K-1}, a_K)|\bar{L}_K, \bar{A}_{K-1}]$$
$$g_{m-1}^{opt}(\bar{L}_{m-1}, \bar{A}_{m-2}) = argmax_{a_{m-1}} \sum_{j>m} \tau_j E[Y_j(\bar{A}_{m-2}, a_{m-1}, \underline{g}_m^{opt})|\bar{L}_{m-1}, \bar{A}_{m-2}].$$

That is, the optimal treatment rule at each time step is the rule that maximizes the weighted sum of expected future counterfactual outcomes assuming that the optimal treatment rule is followed at all future time steps. Hence, in terms of the blip function,

$$g_m^{opt}(\bar{L}_m, \bar{A}_{m-1}) = argmax_{a_m} \sum_{j>m} \tau_j \gamma_{mj}^{g^{opt}*}(\bar{L}_m, \bar{A}_{m-1}, a_m). \tag{37}$$

We can specify a parametric model $\gamma_{mj}^{g^{opt}}(\bar{L}_m, \bar{A}_{m-1}, a_m) = \gamma_{mj}^{g^{opt}}(\bar{L}_m, \bar{A}_{m-1}, a_m; \psi_{g^{opt}})$ and estimate $\psi_{g^{opt}}$ via g-estimation as follows. First, plug g^{opt} into (33) to obtain

$$H_{mk}(\gamma^{g^{opt}}) \equiv Y_k + \sum_{j=m}^{k} \{\gamma_{jk}^{g^{opt}}(\bar{L}_j, \bar{A}_{j-1}, argmax_{a_j} \sum_{r=j}^{K} \tau_r \gamma_{jr}^{g^{opt}}(\bar{L}_j, \bar{A}_{j-1}, a_j)) - \gamma_{jk}^{g^{opt}}(\bar{L}_j, \bar{A}_{j-1}, A_j)\}.$$

Plugging $H_{mk}(\gamma^{g^{opt}})$ into the estimating functions of Theorem 3 in place of $H_{mk}(\gamma)$ shows identifiability of $\gamma^{g^{opt}*}$. And by solving the estimating equations in Theorem 4 with $H_{mk}(\psi_{g^{opt}})$ in place of $H_{mk}(\psi)$, we obtain a $n^{1/2}$ consistent estimator of $\psi^*_{g^{opt}}$. We can obtain a consistent estimate of the optimal treatment rule as

$$\hat{g}^{opt}(\bar{L}_m, \bar{A}_{m-1}) = argmax_{a_m} \sum_{j \geq m}^{K} \tau_j \gamma_{mj}^{g^{opt}}(\bar{L}_m, \bar{A}_{m-1}, a_m; \hat{\psi}_{g^{opt}}). \tag{38}$$

And we obtain a consistent estimate of the expected value of the counterfactual utility $E[Y(g^{opt})]$ under the optimal regime as $\hat{E}[Y(g^{opt})] = \sum_{k=0}^{K} \tau_k P_n[H_{0k}(\hat{\psi}_{g^{opt}})]$.

7 Real Data Applications

We illustrate our approach with two applications to real data. We fit a coarse SNMM to model effects of bank deregulation on housing prices using data previously analyzed by Chaisemartin and D'Haultfoeuille (2021) and Favara and Imbs (2015). We also fit a standard SNMM to model effects of floods on flood insurance take-up using data previously analyzed by Gallagher (2014). Code and data for these analyses can be found at https://github.com/zshahn/did_snmm.

7.1 Impact of Bank Deregulation on Housing Prices

The Interstate Banking and Branching Efficiency Act, passed in 1994, allowed banks to operate across state borders without authorization from states. However, even after the bill was passed every state still imposed certain restrictions on interstate banking. Over time and in a staggered fashion, most states removed at least one of these restrictions. We estimate the effects of initial deregulation treated as a binary indicator variable on housing prices at the county level. Table 1 summarizes treatment timing over the period 1995-2005 considered in the study. To illustrate the benefits of our approach, we fit two coarse SNMMs–one with no time-varying covariates, and one that adjusts for and estimates effect heterogeneity as a function of mortgage volume in the previous year.

Year	# Counties First Deregulating
1995	2
1996	246
1997	119
1998	416
2000	77
2001	55
Never	128

Table 1: Staggered deregulation by county

The no covariate coarse SNMM assumes that parallel trends holds unconditionally, i.e. $E[Y_k(\infty)-Y_{k-1}(\infty)|T=m] = [Y_k(\infty)-Y_{k-1}(\infty)|T>m]$ for all m and k. We specified a nonparametric blip model $E[Y_k(m,1)-Y_k(\infty)|T=m] = \psi^c_{mk}$ with a separate parameter for each effect. We also specified nonparametric nuisance models for $E[A_m|T \geq m]$ and $E[H^c_{mk}(\psi^c) - H^c_{mk-1}(\psi^c)|T \geq m]$ with separate parameters for each m and each m, k, respectively. We obtained point estimates using closed form linear estimators and estimated standard errors via bootstrap.

The coarse SNMM conditioning on mortgage volume from the previous year assumes $E[Y_k(\infty) - Y_{k-1}(\infty)|T = m, \bar{L}_m] = [Y_k(\infty) - Y_{k-1}(\infty)|T = m, \bar{L}_m]$ with L_m denoting log mortgage volume in year $m-1$ and for all m and k. Our blip model $E[Y_k(m,1) - Y_k(\infty)|T = m, \bar{L}_m] = \tau_{mk} + \beta L_m$ posited that the effect varies flexibly with time and linearly with the previous year's log mortgage volume. We specified nuisance models $E[A_m|T \geq m, \bar{L}_m] = \beta_{m0} + \beta_{m1}L_m + \beta_{m2}L^2_m$ and $E[H^c_{mk}(\psi^c) - H^c_{mk-1}(\psi^c)|T \geq m, \bar{L}_m] =$

$\lambda_{mk0} + \lambda_{mk1} L_m$. We again obtained point estimates using closed form linear estimators and estimated standard errors via bootstrap.

Figure 1 displays the estimated effects on log housing price index of the deregulation that actually occurred compared to a counterfactual in which there was no deregulation (i.e. $E[Y_k - Y_k(\infty)]$ for each k). Estimated effects increase over time, which reflects some combination of effects of deregulation at a given location growing over time and increasing numbers of deregulated locations over time. Figure 2 displays for each lag t the average effect over all deregulations t years after the deregulation, i.e. $\frac{1}{\sum_{i=1}^{N} \sum_{m=1995}^{2005} \mathbb{1}\{T_i = m\}} \sum_{i=1}^{N} \sum_{m=1995}^{2005} \mathbb{1}\{T_i = m\} \hat{\gamma}^c_{m,m+t}(\bar{L}_{im})$. The quantity plotted in Figure 2 is also considered by Chaisemartin and D'Haultfoeuille (2021) and illustrates that on average a deregulation's effects are estimated to grow over time. We do not directly compare our results with Chaisemartin and D'Haultfoeuille (2021) or Favara and Imbs (2015) because they each consider somewhat different estimands, but each of those analyses also found that deregulation increased housing prices and that effects grew over time.

In both Figures 1 and 2, the effect estimates from the coarse SNMM conditioning on mortgage volume are qualitatively similar but statistically significantly different from the effect estimates generated by the unconditional coarse SNMM, particularly at earlier years in Figure 1 and shorter time lags in Figure 1. At shorter time lags, the unconditional coarse SNMM estimates small negative effects of deregulation on housing prices, while the coarse SNMM conditional on mortgage volume estimates small positive effects. Perhaps the discrepancy arises because conditioning on mortgage volume corrects some bias.

We can also examine the estimate of the parameter β from conditional coarse SNMM $E[Y_k(m,1) - Y_k(\infty) | T = m, \bar{L}_m] = \tau_{mk} + \beta L_m$ to learn about effect heterogeneity as a function of mortgage volume. We obtained the estimate $\hat{\beta} = 0.044$ $(.95CI = [0.026, 0.061])$, which is evidence that the effect of deregulation on housing prices is much greater in counties with higher mortgage volume. The interquartile range of log mortgage volume was 2.6, corresponding to a difference of 0.11 in estimated average effect of deregulation on housing prices between counties in the third and first quartiles of the mortgage volume distribution. This heterogeneity is quite strong relative to the scale of the average effects in Figures 1 and 2.

7.2 Impact of Floods on Flood Insurance

Gallagher (2014) used a fixed effects regression model to look at effects of floods on flood insurance coverage at the county level. He argued that each county's flood risk is constant over time. We fit a standard SNMM under the assumption of parallel trends in insurance coverage absent future floods in counties with similar flood history from 1958. We specified a parametric linear blip model $\gamma_{mk}(\bar{l}_m, \bar{a}_m) = a_m(1, m - 1980, k - m, (k-m)^2, rate_{m-1})^T \psi$, where $rate_{m-1}$ denotes the county's proportion of flood years since 1958. We specified nuisance models $E[A_m | \bar{A}_{m-1}, \bar{L}_m] = \beta_{m0} + \beta_{m1} rate_{m-1}$ and $E[H_{mk}(\psi) - H_{mk-1}(\psi) | \bar{A}_{m-1}, \bar{L}_m] = \lambda_{mk0} + \lambda_{mk1} rate_{m-1}$. We obtained blip model parameter estimates via closed form linear estimators and estimated standard errors via

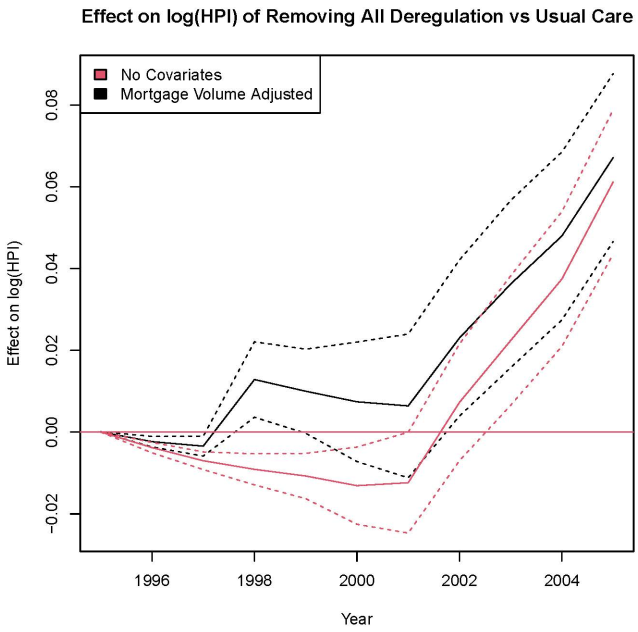

Figure 1: Estimated effects on log housing price index of the deregulation that actually occurred compared to a counterfactual in which there was no deregulation, i.e. $E[Y_k - Y_k(\infty)]$ for each k

bootstrap. Each line in Figure 3 depicts the estimated effect on flood insurance uptake over 15 years of a flood at the leftmost time point on the line followed by no further floods over the 15 year period for a county with the median historical flood rate, i.e. $\gamma_{mk}(A_m = 1, rate_{m-1} = rate_{median}; \hat{\psi})$ for $k \in m, \ldots, m+15$. These quantities were directly extracted from our blip function estimates. We see that there is an initial surge in uptake followed by a steep decline, and the estimated initial surge is larger for more recent floods. We did not find statistically significant effect heterogeneity as a function of historical county flood rate. Gallagher (2014) obtained qualitatively similar results and argued that most of the decline in the effect of a flood is due to residents forgetting about it as opposed to migration. It might be interesting to explore other blip model specifications, perhaps conditioning on further aspects of flood history such as years since previous flood or on average flood insurance premiums in the area.

8 Extensions and Future Work

8.1 Efficiency Theory

In future work, it would be of interest to derive the efficient influence functions of coarse and standard SNMM parameters under parallel trends assumptions to assess

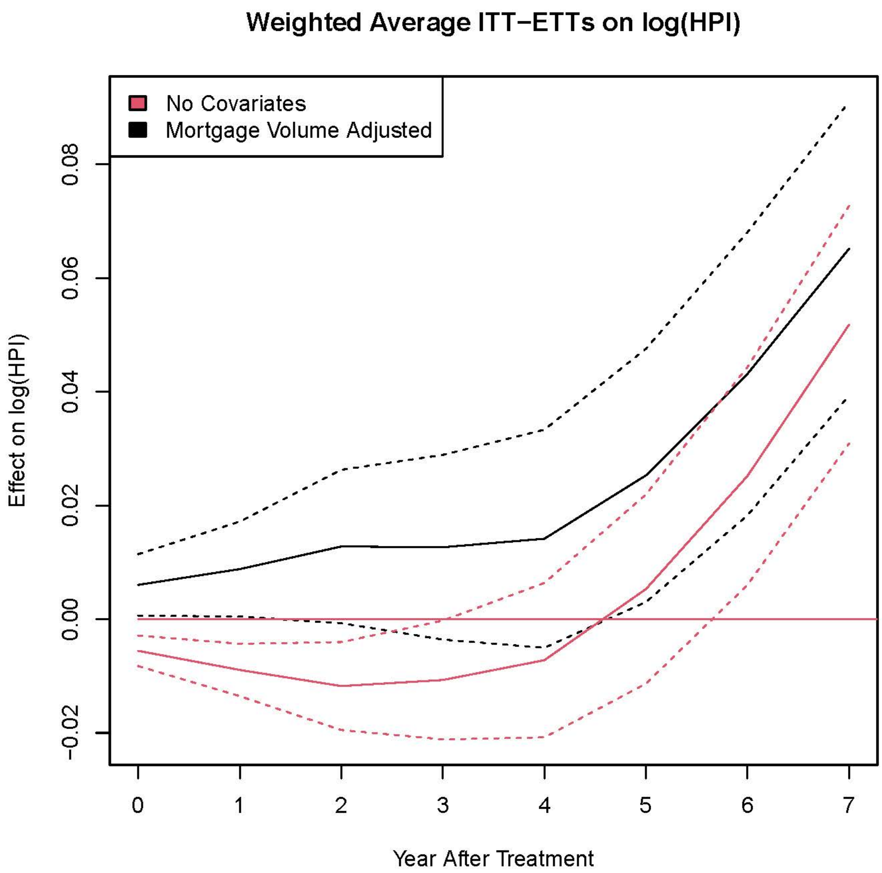

Figure 2: For each lag t the average effect over all deregulations t years after the deregulation, i.e. $\frac{1}{\sum_{i=1}^{N}\sum_{m=1995}^{2005}\mathbb{1}\{T_i=m\}}\sum_{i=1}^{N}\sum_{m=1995}^{2005}\mathbb{1}\{T_i=m\}\hat{\gamma}^{c}_{m,m+t}(\bar{L}_{im})$.

whether our proposed estimators achieve semi-parametric efficiency bounds and, if not, to construct estimators that do. It would further be interesting to compare the relative efficiency of our proposed estimators to those of Callaway and Sant'anna (2021) and Chaisemartin and D'Haultfoeuille (2020) for settings in which they target the same causal estimands. In particular we would like to explore whether our approach enjoys efficiency gains from incorporating time-varying covariates when both our version of the parallel trends assumption (conditioning on time-varying covariates) and the versions of the parallel trends assumption in Callaway and Sant'anna (2021) and Chaisemartin and D'Haultfoeuille (2020) (conditioning only on baseline covariates) hold.

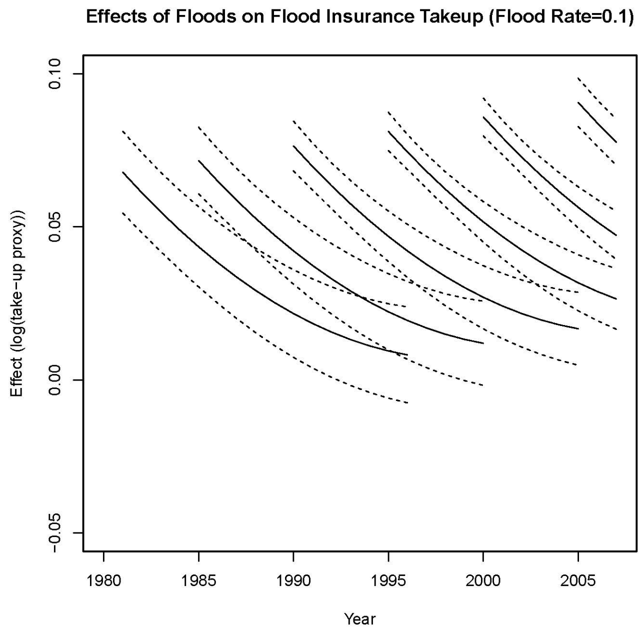

Figure 3: Each line in Figure 3 depicts the estimated effect on flood insurance uptake over 15 years of a flood at the leftmost time point on the line followed by no further floods over the 15 year period for a county with the median historical flood rate, i.e. $\gamma_{mk}(A_m = 1, rate_{m-1} = rate_{median}; \hat{\psi})$ for $k \in m, \ldots, m+15$. These quantities were directly extracted from our blip function estimates.

8.2 Sensitivity Analysis

Conditional parallel trends assumptions are strong and untestable, and sensitivity analysis for violations of the parallel trends assumption is therefore desirable. We adapt the approach to sensitivity analysis for unobserved confounding in SNMMs of Robins et al (2000) and Yang and Lok (2018) to sensitivity analysis for non-parallel trends. We describe a general class of bias functions characterizing deviations from parallel trends given covariate history. For any particular bias function from this class, we provide a corresponding unbiased estimate of (coarse) SNMM parameters assuming that the bias function is correctly specified. An analyst can then execute a sensitivity analysis by specifying a plausible range of bias functions (e.g. a grid of parameters covering a plausible range within a parametric subclass of bias functions) and producing the corresponding range of plausible effect estimates. This approach to sensitivity analysis is complementary to that developed by Rambachan and Roth (2022), as it allows deviations to depend on covariates and allows for sensitivity analysis of all SNMM parameters (e.g. those characterizing effect heterogeneity) and derived quantities. In the exposition below, we focus for simplicity on coarse SNMMs with binary treatments, but extensions to the general case are straightforward.

Define

$$c(\bar{l}_m, k) = \\ E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m = 1] - E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m = 0]. \tag{39}$$

$c(\bar{l}_m, k)$ characterizes the magnitude of deviation from the parallel trends assumption (10). It is a general function in that it allows deviations to depend both on covariate history and time horizon k.

Given a bias function (39), define the bias adjusted version of the 'blipped down' quantity (12) as

$$H^{c,a}_{mk}(\gamma^c) \equiv H^c_{mk}(\gamma^c) - \sum_{j=m}^{k} Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)c(\bar{L}_j, k)\mathbb{1}\{T \geq j\}, \tag{40}$$

where H^c_{mk} is defined in (12).

Lemma 1. *If (39) is correctly specified, then*

$$\begin{aligned} &E[H^{c,a}_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})|\bar{L}_m, T \geq m, A_m] \\ &= E[H^{c,a}_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})|\bar{L}_m, T \geq m]. \end{aligned} \tag{41}$$

Proof. See Appendix E. □

Lemma 1 states that under correct specification of the bias function (39), the conditional expectation of $H^{c,a}_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})$ does not depend on A_m. This is the same crucial property satisfied by $H^c_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})$ in (14) that enabled identification of γ^{c*}. Thus, it follows that identification and estimation of γ^{c*} under bias function (39) may proceed exactly as identification and estimation of γ^{c*} under the parallel trends assumption (10) except substituting $H^{c,a}_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^{c*})$ for $H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^{c*})$. In future work, we will flesh out this sensitivity analysis framework and explore suitable parameterizations of the bias function for tractable and informative sensitivity analysis.

8.3 Survival Outcomes

Piccioto et al. (2012) introduced Structural Nested Cumulative Failure Time Models (SNCFTMs) for time to event outcomes. We can also identify the parameters of these models under a delayed parallel survival curves assumption and a delayed treatment effect assumption. The delayed parallel survival curves assumption states that treated and untreated subjects at time m who have similar observed histories through time m have parallel counterfactual untreated survival probability curves after time $m + \tau$ for some duration τ. The delayed treatment effect assumption states that treatment given at time m has no effect on outcomes until at least $m + \tau$. Delayed treatment effects occur frequently, for example many vaccines have delayed effects and exposures to toxic chemicals often take a long time to lead to cancer diagnosis. Delayed parallel survival curves imply that the impact of unobserved confounders on survival probabilities on a multiplicative scale vanishes in less than τ time steps. This strange combination of assumptions (i.e. that treatments must have delayed effects but unobserved confounders only short term effects) leads us not to present our current results for survival outcomes, but we do wish to point out that extension of DiD to time-to-event settings via structural nested models may still be a promising direction for future work.

8.4 Controlled Direct Effects

We mentioned in passing that SNMM treatments may be multi-dimensional. In particular, we wish to emphasize that this enables estimation of controlled direct effects (CDEs) (Robins and Greenland, 1992). Such effects might be of interest either to make the parallel trends assumption more plausible or to explore mechanisms. For an example of strengthening the parallel trends assumption through controlled direct effects, suppose an investigator is interested in effects of Medicaid expansion on some outcome that is impacted both by Medicaid and the minimum wage. Given that states that expand Medicaid in a given year are also more likely to later increase the minimum wage, future minimum wage increases can lead to violations of the parallel trends assumption in an analysis where Medicaid is the sole treatment. However, by estimating the joint effect of Medicaid expansion and no future minimum wage expansion compared to never expanding Medicaid or increasing the minimum wage (i.e. the controlled direct effect of Medicaid expansion setting minimum wage increases to 0), the particular threat to the parallel trends assumption posed by future minimum wage increases is eliminated. See Robins and Greenland (1992) for a general discussion of controlled direct effects for exploring mechanisms. See Blackwell et al (2022) for an alternative approach to estimating CDEs under parallel trends assumptions in a somewhat different setting. We explain how use standard and coarse SNMMs to estimate CDEs under parallel trends in Appendix F.

9 Conclusion

To summarize, we have shown that structural nested mean models expand the set of causal questions that can be addressed under parallel trends assumptions. In particular, we have shown that coarse and standard additive and multiplicative SNMMs can all be identified under time-varying conditional parallel trends assumptions. Using these models, we can do many things that were not previously possible in DiD studies, such as: characterize effect heterogeneity as a function of time-varying covariates (e.g. mortgage volume in the bank deregulation analysis), identify the effect of one final blip of treatment (as in the flood analysis) and other derived contrasts (such as controlled direct effects), and adjust for time-varying trend confounders (again, illustrated by the mortgage volume analysis). We have also explained how to estimate counterfactual expectations under a general and possibly dynamic regime as long as parallel trends holds with respect to that regime. Under the stronger assumption that unobserved confounders are not effect modifiers on the relevant scale, we have also shown that optimal dynamic treatment regimes are identified via optimal regime SNMMs. We hope these new capabilities can be put to use in a wide variety of applications.

10 Bibliography

Athey, Susan, and Guido W. Imbens. "Identification and inference in nonlinear difference-in-differences models." Econometrica 74, no. 2 (2006): 431-497.

Athey, Susan, and Guido W. Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. No. w24963. National Bureau of Economic Research, 2018.

Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard. "Panel experiments and dynamic causal effects: A finite population perspective." Quantitative Economics 12, no. 4 (2021): 1171-1196.

Callaway, Brantly, and Pedro HC Sant'Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics 225, no. 2 (2021): 200-230.

Card, D. and Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. Tech. rep., National Bureau of Economic Research.

De Chaisemartin, Clement, and Xavier D'Haultfoeuille. "Difference-in-differences estimators of intertemporal treatment effects." Available at SSRN 3731856 (2020).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/debiased machine learning for treatment and structural parameters." (2018): C1-C68.

Lewis, Greg, and Vasilis Syrgkanis. "Double/debiased machine learning for dynamic treatment effects via g-estimation." arXiv preprint arXiv:2002.07285 (2020).

Lok, Judith J., and Victor DeGruttola. "Impact of time to start treatment following infection with application to initiating HAART in HIV?positive patients." Biometrics 68, no. 3 (2012): 745-754.

Picciotto, Sally, et al. "Structural nested cumulative failure time models to estimate the effects of interventions." Journal of the American Statistical Association 107.499 (2012): 886-900.

Rambachan, Ashesh, and Jonathan Roth. A More Credible Approach to Parallel Trends. Working Paper, 2022.

Robins, James M. "Correcting for non-compliance in randomized trials using structural nested mean models." Communications in Statistics-Theory and methods 23.8 (1994): 2379-2412.

Robins J.M., 1997, Causal inference from complex longitudinal data, In: Latent Variable Modeling and Applications to Causality, Lecture Notes in Statistics (120), M. Berkane, Editor. NY: Springer Verlag, 69-117.

Robins, James M. "Correction for noncompliance in equivalence trials." Statistics in medicine 17, no. 3 (1998): 269-302.

Robins, James M. "Marginal structural models versus structural nested models as tools for causal inference." In Statistical models in epidemiology, the environment, and clinical trials, pp. 95-133. Springer, New York, NY, 2000.

Robins, James M. "Optimal structural nested models for optimal sequential decisions." In Proceedings of the second seattle Symposium in Biostatistics, pp. 189-326. Springer, New York, NY, 2004.

Robins, James M., and Sander Greenland. "Identifiability and exchangeability for direct and indirect effects." Epidemiology (1992): 143-155.

Robins, James M., Andrea Rotnitzky, and Daniel O. Scharfstein. "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, New York, NY, 2000.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. "What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature." arXiv preprint arXiv:2201.01194 (2022).

Smucler, Ezequiel, Andrea Rotnitzky, and James M. Robins. "A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts." arXiv preprint arXiv:1904.03737 (2019).

Snow, J. (1855). On the mode of communication of Cholera. John Churchill.

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., and Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. Health Services and Outcomes Research Methodology, 14(4), 166-182.

Sofer, Tamar, David B. Richardson, Elena Colicino, Joel Schwartz, and Eric J. Tchetgen Tchetgen. "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, no. 3 (2016): 348.

Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.

Vansteelandt, Stijn, and Marshall Joffe. "Structural nested models and G-estimation: the partially realized promise." Statistical Science 29, no. 4 (2014): 707-731.

Yang, Shu, and Judith J. Lok. "Sensitivity analysis for unmeasured confounding in coarse structural nested mean models." Statistica Sinica 28, no. 4 (2018): 1703.

Zeldow, Bret, and Laura A. Hatfield. "Confounding and Regression Adjustment in Difference-in-Differences." arXiv preprint arXiv:1911.12185 (2019).

Appendix

Appendix A: Proof of Theorem 1

Part(i)

$E[U^c_{mk}(s_m,\gamma^{c*})] = E[E[U^c_{mk}(s_m,\gamma^{c*})|\bar{L}_m,\bar{A}_m]]$
$= E[$
$\mathbb{1}\{T\geq m\}E[H^c_{m,k}(\gamma^{c*}) - H^c_{m,k-1}(\gamma^{c*})|\bar{L}_m,\bar{A}_{m-1}](s_m(k,\overline{L}_m,\overline{A}_m) - E[s_m(k,\overline{L}_m,\overline{A}_m)|\bar{L}_m,\bar{A}_{m-1}])$
$]$
by (14)
$= E[$
$\mathbb{1}\{T\geq m\}E[H^c_{m,k}(\gamma^{c*}) - H^c_{m,k-1}(\gamma^{c*})|\bar{L}_m,\bar{A}_{m-1}]\times$
$E[s_m(k,\overline{L}_m,\overline{A}_m) - E[s_m(k,\overline{L}_m,\overline{A}_m)|\bar{L}_m,\bar{A}_{m-1}]|\bar{A}_{m-1},\bar{L}_m]$
$]$
by nested expectations
$= 0$

$$(42)$$

The above establishes that the true blip functions are a solution to these equations. The proof of uniqueness follows from the two Lemmas below.

Lemma 2. *Any functions γ^c satisfying (15) for all $s_m(k,\bar{l}_m,\bar{a}_m)$ also satisfies (20), i.e.*

$$E[H_{mk}(\gamma)-H_{mk-1}(\gamma)|\bar{A}_m=\bar{a}_m,\bar{L}_m] = E[H_{mk}(\gamma)-H_{mk-1}(\gamma)|\bar{A}_m=(\bar{a}_{m-1},0),\bar{L}_m]\forall k>m.$$

Proof. Since (15) must hold for all $s_m(k,\bar{l}_m,\bar{a}_m)$, in particular it must hold for

$$s_m(k,\bar{l}_m,\bar{a}_m) = E[H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^c)|\bar{A}_m,\bar{L}_m].$$

Plugging this choice of $s_m(k,\bar{l}_m,\bar{a}_m)$ into U^c_{mk}, we get

$$\begin{aligned}
&E[\{H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^c)\}\times\\
&\{E[H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^c)|\bar{A}_m,\bar{L}_m] - E[H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^c)|\bar{L}_m,\bar{A}_{m-1}]\}] = 0\\
&\Longrightarrow\\
&E[H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^c)|\bar{A}_m,\bar{L}_m] = E[H^c_{mk}(\gamma^c) - H^c_{mk-1}(\gamma^c)|\bar{L}_m,\bar{A}_{m-1}],
\end{aligned}$$

which proves the result. □

Lemma 3. *γ^{c*} are the unique functions satisfying (14)*

Proof. We proceed by induction. Suppose for any k there exist other functions $\gamma^{c'}$ in addition to γ^{c*} satisfying (14), i.e.

$$\begin{aligned}
&E[H^c_{(k-1)k}(\gamma^{c*}) - H^c_{(k-1)(k-1)}(\gamma^{c*})|(\bar{A}_{k-2}=0, A_{k-1}), \bar{L}_{k-1}] = \\
&E[H^c_{(k-1)k}(\gamma^{c*}) - H^c_{(k-1)(k-1)}(\gamma^{c*})|(\bar{A}_{k-2}=0, 0), \bar{L}_{k-1}] \\
&\mathit{and} \\
&E[H^c_{(k-1)k}(\gamma^{c'}) - H^c_{(k-1)(k-1)}(\gamma^{c'})|(\bar{A}_{k-2}=0, 0), \bar{L}_{k-1}] = \\
&E[H^c_{(k-1)k}(\gamma^{'}) - H^c_{(k-1)(k-1)}(\gamma^{c'})|(\bar{A}_{k-2}=0, 0), \bar{L}_{k-1}].
\end{aligned}$$

Differencing both sides of the above equations using the definition of $H^c_{mk}(\gamma^c)$ yields that

$$\begin{aligned}
&E[\gamma^{c*}_{(k-1)k}(\bar{L}_{k-1})\mathbb{1}\{T=k-1\} - \gamma^{c'}_{(k-1)k}(\bar{L}_{k-1})\mathbb{1}\{T=k-1\}|(\bar{A}_{k-2}=0, A_{k-1}), \bar{L}_{k-1}] \\
&= \gamma^{c*}_{(k-1)k}(\bar{L}_{k-1})\mathbb{1}\{T=k-1\} - \gamma^{c'}_{(k-1)k}(\bar{L}_{k-1})\mathbb{1}\{T=k-1\} \\
&= E[\gamma^{c*}_{(k-1)k}(\bar{L}_{k-1})\mathbb{1}\{T=k-1\} - \gamma^{c'}_{(k-1)k}(\bar{L}_{k-1})\mathbb{1}\{T=k-1\}|(\bar{A}_{k-2}=0, 0), \bar{L}_{k-1}] \\
&= 0.
\end{aligned}$$

This establishes that $\gamma^{c*}_{(k-1)k} = \gamma^{c'}_{(k-1)k}$ for all k. Suppose for the purposes of induction that we have established that $\gamma^{c*}_{(k-j)k} = \gamma^{c'}_{(k-j)k}$ for all k and all $j \in \{1, \ldots, n\}$. Now consider $\gamma^{c*}_{(k-(n+1))k}$ and $\gamma^{c'}_{(k-(n+1))k}$. Again we have that

$$\begin{aligned}
&E[H^c_{(k-(n+1))k}(\gamma^{c*}) - H^c_{(k-(n+1))(k-1)}(\gamma^{c*})|(\bar{A}_{k-(n+2)}=0, A_{k-(n+1)}), \bar{L}_{k-(n+1)}] = \\
&E[H^c_{(k-(n+1))k}(\gamma^{c*}) - H^c_{(k-(n+1))(k-1)}(\gamma^{c*})|(\bar{A}_{k-(n+2)}=0, 0), \bar{L}_{k-(n+1)}] \\
&\mathit{and} \\
&E[H^c_{(k-(n+1))k}(\gamma^{c'}) - H^c_{(k-(n+1))(k-1)}(\gamma^{c'})|(\bar{A}_{k-(n+2)}=0, A_{k-(n+1)}), \bar{L}_{k-(n+1)}] = \\
&E[H^c_{(k-(n+1))k}(\gamma^{c'}) - H^c_{(k-(n+1))(k-1)}(\gamma^{c'})|(\bar{A}_{k-(n+2)}=0, 0), \bar{L}_{k-(n+1)}].
\end{aligned}$$

And again we can difference both sides of these equations plugging in the expanded definition of $H^c_{(k-(n+1))k}(\gamma^c)$ to obtain

$$\begin{aligned}
&E[\{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^{c*}_{jk}(\bar{L}_j)\mathbb{1}\{T=j\}\}- \\
&\{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^{c'}_{jk}(\bar{L}_j)\mathbb{1}\{T=j\}\}|(\bar{A}_{k-(n+2)}=0, A_{k-(n+1)}), \bar{L}_{k-(n+1)}] \\
&= E[\{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^{c*}_{jk}(\bar{L}_j)\mathbb{1}\{T=j\}\}- \\
&\{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^{c'}_{jk}(\bar{L}_j)\mathbb{1}\{T=j\}\}|(\bar{A}_{k-(n+2)}, 0), \bar{L}_{k-(n+1)}].
\end{aligned} \tag{43}$$

Under our inductive assumption, (43) reduces to

$$\begin{aligned}
&E[\gamma^{c*}_{(k-(n+1))k}(\bar{L}_{k-(n+1)})\mathbb{1}\{T=k-(n+1)\}- \\
&\gamma^{c'}_{(k-(n+1))k}(\bar{L}_{k-(n+1)})\mathbb{1}\{T=k-(n+1)\}|(\bar{A}_{k-(n+2)}=0, A_{k-(n+1)}), \bar{L}_{k-(n+1)}] \\
&= \gamma^{c*}_{(k-(n+1))k}(\bar{L}_{k-(n+1)})\mathbb{1}\{T=k-(n+1)\} - \gamma^{c'}_{(k-(n+1))k}(\bar{L}_{k-(n+1)})\mathbb{1}\{T=k-(n+1)\} \\
&= E[\gamma^{c*}_{(k-(n+1))k}(\bar{L}_{k-(n+1)})\mathbb{1}\{T=k-(n+1)\}- \\
&\gamma^{c'}_{(k-(n+1))k}(\bar{L}_{k-1})\mathbb{1}\{T=k-(n+1)\}|(\bar{A}_{k-(n+2)}=0, 0), \bar{L}_{k-(n+1)}] \\
&= 0,
\end{aligned}$$

proving that $\gamma^{c*}_{(k-(n+1))k} = \gamma^{c'}_{(k-(n+1))k}$ for all k. Hence, by induction, the result follows. □

Uniqueness is a direct corollary of Lemmas 2 and 3.

Part (ii)

$$
\begin{aligned}
&E[U^{c\dagger}_{mk}(s_m, \underline{\gamma}^{c*}_m, \widetilde{v}^c_m, \widetilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}})] = E[E[U^{c\dagger}_{mk}(s_m, \underline{\gamma}^{c*}_m, \widetilde{v}^c_m, \widetilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}})|\bar{L}_m, \bar{A}_m]] \\
&= E[\\
&E[H^c_{m,k}(\gamma^{c*}) - H^c_{m,k-1}(\gamma^{c*}) - \tilde{v}^c_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^{c*})|\bar{L}_m, \bar{A}_{m-1}]q(\overline{L}_m, \overline{A}_{m-1})\times \\
&(s_m(k, \overline{L}_m, \overline{A}_m) - \tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}]) \\
&] \\
&\text{by (14)} \\
&= E[\\
&E[E[H^c_{m,k}(\gamma^{c*}) - H^c_{m,k-1}(\gamma^{c*}) - \tilde{v}^c_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^{c*})|\bar{L}_m, \bar{A}_{m-1}]q(\overline{L}_m, \overline{A}_{m-1})\times \\
&E[s_m(k, \overline{L}_m, \overline{A}_m) - \tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}]|\bar{L}_m, \bar{A}_{m-1}]] \\
&]
\end{aligned} \tag{44}
$$

Now the result follows because if $\tilde{v}^c_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^{c*}) = v^c_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^{c*})$, then

$$E[E[H^c_{m,k}(\gamma^{c*}) - H^c_{m,k-1}(\gamma^{c*}) - \tilde{v}^c_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^{c*})|\bar{L}_m, \bar{A}_{m-1}] = 0$$

and if $\tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}} = E_{A_m|\overline{L}_m,\overline{A}_{m-1}}$ is correctly specified then

$$E[s_m(k, \overline{L}_m, \overline{A}_m) - \tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}]|\bar{L}_m, \bar{A}_{m-1}]] = 0.$$

Appendix B: Illustrative Description of a Hypothetical Nonadherence Adjustment Using a Coarse SNMM

We illustrate the substantive advantages of our approach when time-varying covariates come into play with an example. Suppose the goal is to adjust for non-adherence in an arm of a randomized trial by estimating the expected counterfactual outcome in that arm under full adherence. To make the example a bit more concrete (if still stylized), suppose the trial is comparing anti-hypertensive medications. Subjects attend K monthly visits where their blood pressure (the outcome of interest, Y_t at month t) and other measurements L_t such as self-reported stress are recorded. We assume all visits are attended and there is no loss to followup. At each visit, subjects also report truthfully whether they were adherent in the previous month (A_{t-1}, 0 indicating adherence and 1 non-adherence).

A coarse SNMM models $\gamma^c_{mk}(\bar{l}_m) = E[Y_k(m) - Y_k(\infty)|\bar{L}_m = \bar{l}_m, T = m]$, which is the effect on blood pressure at month k of stopping medication for the first time at month m compared to never stopping medication among patients who stopped medication for the first time at month m and had observed covariate history $\bar{l}_m$ through month m. Note that subjects may go on and off medication repeatedly, but we can still specify a coarse SNMM that marginalizes over future adherence patterns. Robins (1998) explained how to estimate $\gamma^c_{mk}(\bar{l}_m)$ under a no unobserved confounders assumption and then estimate the expected outcome under full adherence $E[\bar{Y}_K(\infty)]$ as a derived quantity.

But suppose we believe there is unobserved confounding. If we further believe a parallel trends assumption conditional on just baseline covariates such as age and sex, we can estimate $E[Y_k(m)-Y_k(\infty)|L_0, T=m]$, $E[Y_k(m)-Y_k(\infty)|T=m]$, $E[Y_k(\infty)|T=m]$, and $E[\bar{Y}_K(\infty)]$ with either our estimators or previous time-varying DiD methods. But suppose we are concerned that higher stress levels are associated with increased non-adherence and with slower declines in blood pressure over time under full adherence. We might dub stress a time varying 'trend confounder'. In the presence of time-varying trend confounding by stress, parallel trends assumptions unconditional on time-varying stress level would not hold. But our parallel trends assumption (10) conditional on time-varying stress level might hold, allowing us to estimate $E[Y_k(\infty)]$ for each k as $\mathbb{P}[H^c_{0k}(\gamma^c(\hat{\psi}_c))]$. We note that while we showed nonparametric identification in Theorem 1, in practice a non-saturated parametric coarse SNMM model will often need to be correctly specified to enable consistent estimation. Then, in order to adjust for time-varying confounding of trends using our method with a parametric coarse SNMM, we need the additional assumption (compared to previous methods) that the model (9) of effect heterogeneity by time-varying trend confounders is well specified.

Suppose that we are also interested in how the effect of non-adherence varies with stress level at the time of non-adherence. To answer this question, we could specify a coarse SNMM that depends on stress and estimate its parameters using cross-fit estimators from Theorem 2 (or using parametric nuisance models as in Remarks 4 and 5). Suppose for simplicity that our only time-varying covariate L_m is a binary indicator for high stress at month m. We could separately estimate $E[Y_k(m)-Y_k(\infty)|T=m, L_m=1]$ and $E[Y_k(m)-Y_k(\infty)|T=m, L_m=0]$ for each $k>m$ with $\mathbb{P}^{T=m,L_m=1}[H^c_{mk}(\gamma^c(\hat{\psi}^{cf}_c))]$ and $\mathbb{P}^{T=m,L_m=0}[H^c_{mk}(\gamma^c(\hat{\psi}^{cf}_c))]$, respectively. The difference between the estimates conditioning on $L_m=0$ and $L_m=1$ can characterize the importance of stress as a modifier of the effect of going off medication. Alternatively, under certain parametric coarse SNMM model specifications, it might be possible to assess effect heterogeneity by stress by simply examining the parameter estimates $\hat{\psi}^{cf}_c$. To our knowledge, it is not possible to address questions pertaining to time-varying effect heterogeneity at all using previously developed DiD estimators.

Appendix C: Proof of Theorem 3

Part (i)

$E[U_{mk}(s_m, \gamma^*)] = E[E[U_{mk}(s_m, \gamma^*)|\bar{L}_m, \bar{A}_m]]$
$= E[$
$E[H_{m,k}(\gamma^*) - H^c_{m,k-1}(\gamma^*)|\bar{L}_m, \bar{A}_{m-1}](s_m(k, \overline{L}_m, \overline{A}_m) - E[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}])$
$]$
by (20)
$= E[$
$E[H_{m,k}(\gamma^*) - H_{m,k-1}(\gamma^*)|\bar{L}_m, \bar{A}_{m-1}]\times$
$E[s_m(k, \overline{L}_m, \overline{A}_m) - E[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}]|\bar{A}_{m-1}, \bar{L}_m]$
$]$
by nested expectations
$= 0$

(45)

The above establishes that the true blip functions are a solution to these equations. The proof of uniqueness follows from the two Lemmas below.

Lemma 4. *Any functions γ satisfying (21) for all $s_m(k, \bar{l}_m, \bar{a}_m)$ also satisfies (20), i.e.*

$$E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{A}_m = \bar{a}_m, \bar{L}_m] = E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{A}_m = (\bar{a}_{m-1}, 0), \bar{L}_m]\forall k > m.$$

Proof. Since (21) must hold for all $s_m(k, \bar{l}_m, \bar{a}_m)$, in particular it must hold for

$$s_m(k, \bar{l}_m, \bar{a}_m) = E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{A}_m, \bar{L}_m].$$

Plugging this choice of $s_m(k, \bar{l}_m, \bar{a}_m)$ into U_{mk}, we get

$$E[\{H_{mk}(\gamma) - H_{mk-1}(\gamma)\}\times$$
$$\{E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{A}_m, \bar{L}_m] - E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{L}_m, \bar{A}_{m-1}]\}] = 0$$
$$\Longrightarrow$$
$$E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{A}_m, \bar{L}_m] = E[H_{mk}(\gamma) - H_{mk-1}(\gamma)|\bar{L}_m, \bar{A}_{m-1}],$$

which proves the result. □

Lemma 5. *γ^* are the unique functions satisfying (20)*

Proof. We proceed by induction. Suppose for any k there exist other functions $\gamma^{'}$ in addition to γ^* satisfying (20) and satisfying $\gamma^{'}_{mk}(\bar{l}_m, (\bar{a}_{m-1}, 0)) = 0$. Then

$$E[H_{(k-1)k}(\gamma^*) - H_{(k-1)(k-1)}(\gamma^*)|\bar{A}_{k-1}, \bar{L}_{k-1}] = E[H_{(k-1)k}(\gamma^*) - H_{(k-1)(k-1)}(\gamma^*)|(\bar{A}_{k-2}, 0), \bar{L}_{k-1}]$$
$$E[H_{(k-1)k}(\gamma^{'}) - H_{(k-1)(k-1)}(\gamma^{'})|\bar{A}_{k-1}, \bar{L}_{k-1}] = E[H_{(k-1)k}(\gamma^{'}) - H_{(k-1)(k-1)}(\gamma^{'})|(\bar{A}_{k-2}, 0), \bar{L}_{k-1}].$$

Differencing both sides of the above equations using the definition of $H_{mk}(\gamma)$ yields that

$$
\begin{aligned}
&E[\gamma^*_{(k-1)k}(\bar{L}_{k-1}, \bar{A}_{k-1}) - \gamma^{'}_{(k-1)k}(\bar{L}_{k-1}, \bar{A}_{k-1})|\bar{A}_{k-1}, \bar{L}_{k-1}] \\
&= \gamma^*_{(k-1)k}(\bar{L}_{k-1}, \bar{A}_{k-1}) - \gamma^{'}_{(k-1)k}(\bar{L}_{k-1}, \bar{A}_{k-1}) \\
&= E[\gamma^*_{(k-1)k}(\bar{L}_{k-1}, \bar{A}_{k-1}) - \gamma^{'}_{(k-1)k}(\bar{L}_{k-1}, \bar{A}_{k-1})|(\bar{A}_{k-2}, 0), \bar{L}_{k-1}] \\
&= 0,
\end{aligned}
$$

where the last equality follows from the assumption that $\gamma^{'}_{mk}(\bar{a}_m, \bar{l}_m) = 0$ when $a_m = 0$. This establishes that $\gamma^*_{(k-1)k} = \gamma^{'}_{(k-1)k}$ for all k. Suppose for the purposes of induction that we have established that $\gamma^*_{(k-j)k} = \gamma^{'}_{(k-j)k}$ for all k and all $j \in \{1, \dots, n\}$. Now consider $\gamma^*_{(k-(n+1))k}$ and $\gamma^{'}_{(k-(n+1))k}$. Again we have that

$$
\begin{aligned}
&E[H_{(k-(n+1))k}(\gamma^*) - H_{(k-(n+1))(k-1)}(\gamma^*)|\bar{A}_{k-(n+1)}, \bar{L}_{k-(n+1)}] = \\
&E[H_{(k-(n+1))k}(\gamma^*) - H_{(k-(n+1))(k-1)}(\gamma^*)|(\bar{A}_{k-(n+2)}, 0), \bar{L}_{k-(n+1)}] \\
&\textit{and} \\
&E[H_{(k-(n+1))k}(\gamma^{'}) - H_{(k-(n+1))(k-1)}(\gamma^{'})|\bar{A}_{k-(n+1)}, \bar{L}_{k-(n+1)}] = \\
&E[H_{(k-(n+1))k}(\gamma^{'}) - H_{(k-(n+1))(k-1)}(\gamma^{'})|(\bar{A}_{k-(n+2)}, 0), \bar{L}_{k-(n+1)}].
\end{aligned}
$$

And again we can difference both sides of these equations plugging in the expanded definition of $H_{(k-(n+1))k}(\gamma)$ to obtain

$$
\begin{aligned}
&E[\{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^*_{jk}(\bar{A}_j, \bar{L}_j)\} - \{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^{'}_{jk}(\bar{A}_j, \bar{L}_j)\}|\bar{A}_{k-(n+1)}, \bar{L}_{k-(n+1)}] \\
&= E[\{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^*_{jk}(\bar{A}_j, \bar{L}_j)\} - \{Y_k - \sum_{j=k-(n+1)}^{k-1} \gamma^{'}_{jk}(\bar{A}_j, \bar{L}_j)\}|(\bar{A}_{k-(n+2)}, 0), \bar{L}_{k-(n+1)}].
\end{aligned}
\tag{46}
$$

Under our inductive assumption, (46) reduces to

$$
\begin{aligned}
&E[\gamma^*_{(k-(n+1))k}(\bar{L}_{k-(n+1)}, \bar{A}_{k-(n+1)}) - \gamma^{'}_{(k-(n+1))k}(\bar{L}_{k-(n+1)}, \bar{A}_{k-(n+1)})|\bar{A}_{k-(n+1)}, \bar{L}_{k-(n+1)}] \\
&= \gamma^*_{(k-(n+1))k}(\bar{L}_{k-(n+1)}, \bar{A}_{k-(n+1)}) - \gamma^{'}_{(k-(n+1))k}(\bar{L}_{k-(n+1)}, \bar{A}_{k-(n+1)}) \\
&= E[\gamma^*_{(k-(n+1))k}(\bar{L}_{k-(n+1)}, \bar{A}_{k-(n+1)}) - \gamma^{'}_{(k-(n+1))k}(\bar{L}_{k-1}, \bar{A}_{k-(n+1)})|(\bar{A}_{k-(n+2)}, 0), \bar{L}_{k-(n+1)}] \\
&= 0,
\end{aligned}
$$

proving that $\gamma^*_{(k-(n+1))k} = \gamma^{'}_{(k-(n+1))k}$ for all k. Hence, by induction, the result follows. $\square$

Uniqueness is a direct corollary of Lemmas 4 and 5.

Part (ii)

$$E[U^{\dagger}_{mk}(s_m, \underline{\gamma}^*_m, \widetilde{v}_m, \widetilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}})] = E[E[U^{\dagger}_{mk}(s_m, \underline{\gamma}^*_m, \widetilde{v}_m, \widetilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}})|\bar{L}_m, \bar{A}_m]]$$
$$= E[$$
$$E[H_{m,k}(\gamma^{c*}) - H_{m,k-1}(\gamma^*) - \tilde{v}_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^*)|\bar{L}_m, \bar{A}_{m-1}]q(\overline{L}_m, \overline{A}_{m-1})\times$$
$$(s_m(k, \overline{L}_m, \overline{A}_m) - \tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}])$$
$$]$$

by (14)

$$= E[$$
$$E[E[H_{m,k}(\gamma^*) - H_{m,k-1}(\gamma^*) - \tilde{v}_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^*)|\bar{L}_m, \bar{A}_{m-1}]q(\overline{L}_m, \overline{A}_{m-1})\times$$
$$E[s_m(k, \overline{L}_m, \overline{A}_m) - \tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}]|\bar{L}_m, \bar{A}_{m-1}]]$$
$$] \tag{47}$$

Now the result follows because if $\tilde{v}_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^*) = v_m(k, \bar{L}_m, \bar{A}_{m-1}, \gamma^*)$, then

$$E[E[H_{m,k}(\gamma^*) - H_{m,k-1}(\gamma^*) - \tilde{v}^{(}_m k, \bar{L}_m, \bar{A}_{m-1}, \gamma^*)|\bar{L}_m, \bar{A}_{m-1}] = 0$$

and if $\tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}} = E_{A_m|\overline{L}_m,\overline{A}_{m-1}}$ is correctly specified then

$$E[s_m(k, \overline{L}_m, \overline{A}_m) - \tilde{E}_{A_m|\overline{L}_m,\overline{A}_{m-1}}[s_m(k, \overline{L}_m, \overline{A}_m)|\bar{L}_m, \bar{A}_{m-1}]|\bar{L}_m, \bar{A}_{m-1}]] = 0.$$

Appendix D: Relationship Between Universal Parallel Trends and No Additive Effect Modification by Unobserved Confounders Assumptions

To identify the counterfactual expectation $E[Y_k(g)]$ for a given g, we assumed parallel trends under g. We argued, however, that if one assumes that parallel trends happen to hold for a particular regime of interest, one is really in effect assuming that parallel trends holds for *all* regimes $g \in \mathcal{G}$. To identify optimal treatment strategies via optimal regime SNMMs, we needed to make the additional assumption (36) that there is no additive effect modification by unobserved confounders. Below, we sketch an argument that parallel trends under all regimes *effectively*, though not strictly mathematically, implies no effect modification by unobserved confounders.

We will attempt to show the contrapositive of our result of interest, i.e. we will try to show that if there is additive effect modification by an unobserved confounder then parallel trends under all regimes cannot hold. Suppose that $\bar{L}, \bar{U}$ is a minimal set satisfying sequential exchangeability. And suppose that $\bar{U}$ *is* an effect modifier, i.e. for some m, k, and g

$$E[Y_k(\bar{a}_{m-1}, \underline{g}_m) - Y_k(\bar{a}_{m-1}, 0, \underline{g}_{m+1})|\bar{A}_m, \bar{L}_m, \bar{U}_m = \bar{u}_m] \tag{48}$$

depends on $\bar{u}_m$. If parallel trends does hold for all regimes, in particular it would hold for $\underline{g}_m$ and $(0, \underline{g}_{m+1})$, i.e.

$$E[Y_k(\bar{a}_{m-1}, \underline{g}_m) - Y_{k-1}(\bar{a}_{m-1}, \underline{g}_m)|\bar{A}_m, \bar{L}_m] \tag{49}$$

would not depend on A_m and

$$E[Y_k(\bar{a}_{m-1}, 0, \underline{g}_{m+1}) - Y_{k-1}(\bar{a}_{m-1}, 0, \underline{g}_{m+1})|\bar{A}_m, \bar{L}_m] \tag{50}$$

would not depend on A_m. Therefore, the difference of the above two quantities also would not depend on A_m, i.e.

$$E[Y_k(\bar{a}_{m-1}, \underline{g}_m) - Y_k(\bar{a}_{m-1}, 0, \underline{g}_{m+1})|\bar{A}_m, \bar{L}_m] - E[Y_{k-1}(\bar{a}_{m-1}, \underline{g}_m) - Y_{k-1}(\bar{a}_{m-1}, 0, \underline{g}_{m+1})|\bar{A}_m, \bar{L}_m] \tag{51}$$

would not depend on A_m. But U_m being an effect modifier (i.e. (48) depending on u_m) and the assumption that $(\bar{L}_m, \bar{U}_m)$ is a minimal sufficient set together imply that the first conditional expectation in (51) *does* depend on A_m. Now, it is possible for the full difference in (51) to still not depend on A_m if the second conditional expection also depends on A_m and in such a way as to cancel out the dependence on A_m of the first conditional expectation. While this is technically possible, the type of cancelling out required is not plausible. So by reductio ad absurdity (though not the formal Latin absurdum indicating a true contradiction), parallel trends for all regimes "effectively" implies no additive effect modification by unobserved confounders.

Appendix E: Proof of Lemma 1, enabling sensitivity analysis We closely follow Yang and Lok (2018). We first show that the following equalities hold:

$$\begin{aligned}
&E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m] - E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m] \\
&= Pr(1 - A_m|\bar{L}_m, T \geq m)(2A_m - 1)c(\bar{L}_m, k) \\
&= Pr(1 - A_m|\bar{L}_m, T \geq m)\times \\
&\{E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m] - E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, 1 - A_m]\}
\end{aligned} \tag{52}$$

Consider the case where $A_m = 0$. Then

$$\begin{aligned}
&LHS = \\
&E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m] - E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T > m]Pr(A_m = 0|\bar{L}_m, T \geq m) \\
&- E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T = m]Pr(A_m = 1|\bar{L}_m, T \geq m) \\
&= Pr(A_m = 1|\bar{L}_m, T \geq m)\{E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m] - \\
&E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, 1 - A_m]\} \\
&= RHS
\end{aligned} \tag{53}$$

The argument is similar if $A_m = 1$.

For $j > m$, we have that

$$E[Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T = m] = 0 \tag{54}$$

(because $\mathbb{1}\{T \geq j\}$ is 0 under the conditioning event) and

$$\begin{aligned}
&E[Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T > m] \\
&= E[E[Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)|\bar{L}_j, T \geq j]\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T > m] \\
&= E[E[Pr(A_j = 1|\bar{L}_j, T \geq j)Pr(A_j = 0|\bar{L}_j, T \geq j) + Pr(A_j = 1|\bar{L}_j, T \geq j)Pr(A_j = 0|\bar{L}_j, T \geq j) \times (-1)]\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T > m] \\
&= 0.
\end{aligned} \tag{55}$$

Therefore, for $j > m$,

$$\begin{aligned}
&E[Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T \geq m, A_m] \\
&= E[Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T \geq m] \\
&= 0.
\end{aligned} \tag{56}$$

Now,

$$\begin{aligned}
&E[H^{c,a}_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})|\bar{L}_m, T \geq m, A_m] \\
&= E[H^c_{mk}(\gamma^{c*}) - H^c_{mk-1}(\gamma^{c*})|\bar{L}_m, T \geq m, A_m] - Pr(1 - A_m|\bar{L}_m, T \geq m)(2A_m - 1)c(\bar{L}_m, k) \\
&- E[\sum_{j=m+1}^{k-1} Pr(1 - A_j|\bar{L}_j, T \geq j)(2A_j - 1)\mathbb{1}\{T \geq j\}c(\bar{L}_j, k)|\bar{L}_m, T \geq m, A_m] \\
&= E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m] - \\
&\{E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m, A_m] - E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m]\} \\
&= E[Y_k(\infty) - Y_{k-1}(\infty)|\bar{L}_m, T \geq m],
\end{aligned} \tag{57}$$

where: the first equality follows from the definition of $H^{c,a}_{mk}(\gamma^c)$ (40); the second equality follows from (13) (which does not depend on parallel trends), (52), and (56) to make the sum term vanish; and, proving the Lemma, the final term does not depend on A_m.

Appendix F: Estimating Controlled Direct Effects Via SNMMs Under Parallel Trends For a two dimensional treatment $(\bar{A}, \bar{R})$, consider the effects

$$\begin{aligned}
&\gamma^{CDE*}_{mk}((\bar{a}_m, \bar{r}_m), \bar{l}_m) = \\
&E[Y_k(\bar{a} = (\bar{A}_m, \underline{0}), \bar{r} = (\bar{R}_m, \underline{0})) - Y_k(\bar{a} = (\bar{A}_{m-1}, \underline{0}), \bar{r} = (\bar{R}_{m-1}, \underline{0}))|\bar{A}_m = \bar{a}_m, \bar{R}_m = \bar{r}_m, \bar{L}_m = \bar{l}_m].
\end{aligned} \tag{58}$$

These are just the blip functions of a standard SNMM with a two dimensional treatment, and are therefore identified and can be consistently and asymptotically normally estimated by the results of Section 4. Then the quantity $\gamma^{CDE*}_{mk}((\bar{a}_m, \bar{r}_m), \bar{l}_m)$

with $r_m = 0$ might be of particular interest as the controlled direct effect of a treatment a_m at time m setting $\underline{R}_m$ to 0 in subjects with $R_m = 0$ and $\bar{L}_m = \bar{l}_m$. Revisiting the flood example, suppose that $\bar{A}$ denotes flood history and $\bar{R}$ denotes wildfire history. Suppose floods are associated with future wild fires, which is a threat to the validity of the parallel trends assumption in an analysis of floods alone, but interest still centers on the impact of floods. Or, alternatively, suppose that floods make future wildfires less likely by clearing brush or dampening the ground (a possibility that might only be plausible to those as ignorant of this topic as the authors), and interest centers on the *psychological* impact of floods on insurance take-up as opposed to the influence through occurrence of other natural disasters. In either of the above two scenarios, the quantities $\gamma_{mk}^{CDE*}((\bar{a}_m, \bar{r}_m), \bar{l}_m)$ with $r_m = 0$ (perhaps marginalized over $\bar{L}_m$) yielding the controlled direct effects of floods in the absence of future fires would be of interest.

Now, consider the controlled direct effects of a two dimensional coarse intervention

$$E[Y_k(T_A = m, A_m = a_m, T_R = \infty) - Y_k(T_A = T_R = \infty) | T_A = m, A_m = a_m, T_R > m, \bar{L}_m], \tag{59}$$

where T_A denotes the first time of departure of treatment A from its baseline level 0 and T_R denotes the first time of departure of treatment R from its baseline level 0. These are the effects of departing from the A baseline treatment level for the first time at time m and never departing from the R baseline treatment level compared to never departing from either the A or R baseline levels in those who did depart from baseline A treatment at time m and had not yet departed from the baseline R treatment. We have seen how to identify and estimate the second term in the difference (59), i.e. $E[Y_k(T_A = T_R = \infty) | T_A = m, A_m = a_m, T_R > m, \bar{L}_m]$, via a coarse SNMM for two dimensional treatment $\bar{A}, \bar{R}$ in Section 3. To identify and estimate the first term $E[Y_k(T_A = m, A_m = a_m, T_R = \infty) | T_A = m, A_m = a_m, T_R > m, \bar{L}_m]$, we can employ a separate coarse SNMM for effects of R. We will need the R-specific parallel trends assumption

$$\begin{aligned} &E[Y_k(T_A = m, A_m = a_m, T_R = \infty) - Y_{k-1}(T_A = m, A_m = a_m, T_R = \infty) | T_A = m, A_m, T_R \geq j, R_j, \bar{L}_j] = \\ &E[Y_k(T_A = m, A_m = a_m, T_R = \infty) - Y_{k-1}(T_A = m, A_m = a_m, T_R = \infty) | T_A = m, A_m, T_R > j, \bar{L}_j] \\ &\text{for all k>j} \geq \text{m.} \end{aligned} \tag{60}$$

Then, by the results of Section 3, the coarse SNMM for the effects of R in the cohort of subjects with $T_A = m$, i.e.

$$\begin{aligned} &\gamma_{mjk}^{R,c}(\bar{l}_j, r_j, a_m) = \\ &E[Y_k(T_A = m, a_m, T_R = j, r_j) - Y_k(T_A = m, a_m, T_R = \infty) | T_A = m, T_R = j, R_j = r_j, \bar{L}_j] \end{aligned} \tag{61}$$

is identified. We can therefore consistently estimate $E[Y_k(T_A = m, A_m = a_m, T_R = \infty)|T_A = m, A_m = a_m, T_R > m, \bar{L}_m]$ by $\mathbb{P}^{T_A=m,T_R>m}[H^{R,c}_{mk}(\hat{\gamma}^{R,c})]$, where

$$H^{R,c}_{mk}(\gamma^{R,c}) \equiv Y_k - \sum_{j=m}^{k-1} \mathbb{1}\{T_R = j\}\gamma^{R,c}_{mjk}(\bar{L}_j, R_j, a_m). \tag{62}$$

Thus, we can consistently estimate both terms in the difference (59) defining the CDE.