JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 1865–1880: Theory and Methods Special Section on Synthetic Control Methods
https://doi.org/10.1080/01621459.2021.1979561

Prediction Intervals for Synthetic Control Methods

Matias D. Cattaneo[a], Yingjie Feng[b], and Rocio Titiunik[c]

[a]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ; [b]School of Economics and Management, Tsinghua University, Beijing, China; [c]Department of Politics, Princeton University, Princeton, NJ

ABSTRACT

Uncertainty quantification is a fundamental problem in the analysis and interpretation of synthetic control (SC) methods. We develop conditional prediction intervals in the SC framework, and provide conditions under which these intervals offer finite-sample probability guarantees. Our method allows for covariate adjustment and nonstationary data. The construction begins by noting that the statistical uncertainty of the SC prediction is governed by two distinct sources of randomness: one coming from the construction of the (likely misspecified) SC weights in the pretreatment period, and the other coming from the unobservable stochastic error in the post-treatment period when the treatment effect is analyzed. Accordingly, our proposed prediction intervals are constructed taking into account both sources of randomness. For implementation, we propose a simulation-based approach along with finite-sample-based probability bound arguments, naturally leading to principled sensitivity analysis methods. We illustrate the numerical performance of our methods using empirical applications and a small simulation study. Python, R and Stata software packages implementing our methodology are available. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2019
Accepted August 2021

KEYWORDS
Causal inference;
Nonasymptotic inference;
Prediction intervals;
Synthetic controls

1. Introduction

The synthetic control (SC) method was first introduced by Abadie and Gardeazabal (2003) as an approach to study the causal effect of a treatment affecting a single aggregate unit that is observed both before and after the treatment occurs. The authors originally motivated the method with a study of the effect of terrorism in the Basque Country on its GDP per capita. The Basque Country was one of the three richest regions in Spain before the outset of terrorism around the mid-1970s, but the region became relatively poorer in the decades that followed. The question is whether this relative decline can be attributed to terrorism. Their analysis covers the 1955–2000 period and places the beginning of intense terrorism in 1975, thus defining a "pretreatment" period when terrorism is not salient (roughly 1955–1975), and a "post-treatment" period that starts when terrorism intensifies (roughly 1975 onward). The time series data allow for a comparison of Basque GDP before and after the onset of terrorism, but to interpret this change as the causal effect of terrorism would require assuming the absence of time trends. Instead, Abadie and Gardeazabal (2003) proposed to use other regions in Spain, whose GDP is also observed before and after the onset of terrorism in the Basque Country, to build an aggregate or "synthetic" control unit that captures the GDP trajectory that would have occurred in the Basque Country if terrorism had never occurred. The SC is built as a weighted average of all units in the control group (the "donor pool"), where the weights are chosen so that the SC's outcome in the pretreatment period closely matches the treated unit's trajectory while also satisfying some constraints such as being nonnegative, adding up to one, and accounting for other pretreatment covariates. For a contemporaneous review of this literature, see Abadie (2021) and the references therein.

The SC method has received increasing attention since its introduction, and is now a popular component of the methodological toolkit for causal inference and program evaluation (Abadie and Cattaneo 2018). Methodological and theoretical research concerning SC methods has mostly focused either on expanding the SC causal framework (e.g., to dissagregated data or staggered treatment adoption settings) or on developing new implementations of the SC prediction (e.g., via different penalization constraints or matrix completion methods). Recent examples include Abadie and L'Hour (2021), Agarwal et al. (2021), Athey et al. (2021), Bai and Ng (2021), Ben-Michael, Feller, and Rothstein (2021), Chernozhukov, Wüthrich, and Zhu (2021c), Ferman (2021), Kellogg et al. (2021), and Masini and Medeiros (2021); see their references for many more. In contrast, considerably less effort has been devoted to develop principled statistical inference procedures for uncertainty quantification within the SC framework. In particular, Abadie, Diamond, and Hainmueller (2010) proposed a design-based permutation approach under additional assumptions, Li (2020) relied on large-sample approximations for disaggregated data under correct specification, Chernozhukov, Wüthrich, and Zhu (2021b) developed time-series permutation-based inference methods, and Shaikh and Toulis (2021) discussed cross-sectional permutation-based inference methods in semiparametric duration-type settings. See also Feng (2021), and references therein, for related large sample inference

CONTACT Matias D. Cattaneo cattaneo@princeton.edu Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

methods employing local principal component analysis based on nearest-neighbor approximations in possibly nonlinear factor model settings.

We develop conditional prediction intervals for the SC framework, offering an alternative (conditional) inference method to assess statistical uncertainty. Our proposed approach builds on ideas from the literature on conditional prediction intervals (Vovk 2012; Chernozhukov, Wüthrich, and Zhu 2021a) and non-asymptotic concentration (Vershynin 2018; Wainwright 2019) in probability and statistics. As a consequence, the resulting (conditional) prediction intervals are conservative but formally shown to offer probability guarantees. We focus on uncertainty quantification via (conditional) prediction intervals because, in the SC framework, the treatment effect estimator is a random variable emerging from an out-of-sample prediction problem, based on the estimated SC weights constructed using pretreatment data. Our inference procedures are not confidence intervals in the usual sense (i.e., giving a region in the parameter space for a nonrandom parameter of interest), but rather intervals describing a region on the support of a random variable where a new realization is likely to be observed.

Our construction begins by noting that the statistical uncertainty of the SC prediction is governed by two distinct sources of randomness: one due to the construction of the (likely misspecified) SC weights in the pretreatment period, and the other due to the unobservable stochastic error in the post-treatment period when the treatment effect is analyzed. Accordingly, our proposed prediction intervals are constructed taking into account both sources of randomness. For the first source of uncertainty, we propose a simulation-based approach that is justified via non-asymptotic probability concentration and hence enjoys probability guarantees. This approach takes into account the specific construction of the SC weights. For the second source of uncertainty, which comes from out-of-sample prediction due to the unobservable error in the post-treatment period, we discuss several approaches based on nonparametric and parametric probability approximations as a framework for principled sensitivity analysis. This second uncertainty source is harder to handle nonparametrically, and hence its contribution to the overall prediction intervals should be considered with care. Our approach in this article is to employ an agnostic sensitivity analysis, but future work will consider other approaches.

Our results are obtained under high-level conditions, but we provide primitive conditions for three examples: an outcomes-only setting with iid data, a multi-equation setting allowing for stationary weakly dependent data where the weights are obtained by not only matching the pretreatment trends of the outcome of interest but also approximating the trajectories of additional variables such as important covariates or secondary outcomes; and a nonstationary cointegration setting. All three settings allow the weights to be covariate-adjusted in each equation. We also showcase our methods numerically, using both simulated and real data. The methods perform well in finite samples.

The rest of the article proceeds as follows. Section 2 provides a formal introduction to the SC framework and defines the basic quantities of interest. Section 3 introduces the prediction intervals we focus on, and provides basic intuition for their decomposition in terms of the SC weights estimation error and the unobservable post-treatment error. Section 4 develops a simulation-based method to account for the first source of uncertainty, and Section 5 discusses how to (model and) account for the second source of uncertainty. Section 6 illustrates the performance of our proposed prediction intervals with a Monte Carlo experiment and two empirical examples from the SC literature. Section 7 concludes. Appendix A provides an extension of our main in-sample uncertainty quantification approach to the case of weakly dependent (β-mixing) stationary time series data. All the proofs of our technical results, as well as additional numerical evidence, are collected in the online supplemental appendix. We provide companion replication codes in R, and a general-purpose software package is underway (Cattaneo et al. 2021).

2. Setup

We consider the standard SC framework with a single treated unit and several control units, allowing for both stationary and non-stationary data. The data may include only the outcome of interest, or the outcome of interest plus other variables. The researcher observes $N+1$ units for T_0+T_1 periods of time. Units are indexed by $i = 1, 2, \dots N, N+1$, and time periods are indexed by $t = 1, 2, \dots, T_0, T_0+1, \dots, T_0+T_1$. During the first T_0 periods, all units are untreated. Starting at T_0+1, unit 1 receives treatment but the other units remain untreated. Once the treatment is assigned at T_0+1, there is no change in treatment status: the treated unit continues to be treated and the untreated units remain untreated until the end of the series, T_1 periods later.

Each unit i at period t has two potential outcomes, $Y_{it}(1)$ and $Y_{it}(0)$, respectively denoting the outcome under treatment and the outcome in the absence of treatment (which we call the *control* or the *untreated* condition). This notation imposes two additional implicit assumptions that are standard in this setting: no spillovers (the potential outcomes of unit i depend only on i's treatment status) and no anticipation (the potential outcomes at t depend only on the treatment status of the same period).

Attention is restricted to the impact of the treatment on the treated unit. By treatment impact, we mean the difference between the outcome path taken by the treated unit, and the path it would have taken in the absence of the treatment. The quantity of interest is

$$\tau_t = Y_{1t}(1) - Y_{1t}(0), \qquad t > T_0, \tag{1}$$

where τ_t may be regarded as random or nonrandom depending on the framework considered. In this paper, we view τ_t as a random variable.

For each unit, we only observe the potential outcome corresponding to the treatment status actually received by the unit. We denote the observed outcome by Y_{it}, which is defined as follows:

$$Y_{it} = \begin{cases} Y_{it}(0) & \text{if } i = 2, \dots N+1 \\ Y_{it}(0) & \text{if } i = 1 \text{ and } t \in \{1, 2, \dots, T_0\} \\ Y_{it}(1) & \text{if } i = 1 \text{ and } t \in \{T_0+1, \dots, T_0+T_1\} \end{cases}.$$

This means that, in τ_t, the treated unit's potential outcome $Y_{1t}(0)$ is unobservable for all $t > T_0$. The idea of the SC method

is to use an appropriate combination of the post-treatment observed outcomes of the untreated units to approximate the treated unit's counterfactual post-treatment outcome, $Y_{1t}(0)$ for $t > T_0$. This idea has been formalized in different ways since it was originally proposed by Abadie and Gardeazabal (2003).

In all SC frameworks, the formalization chooses a set of weights $\mathbf{w} = (w_2, w_3, \ldots, w_{N+1})'$ such that a given loss function is minimized under constraints. Given a set of estimated weights $\widehat{\mathbf{w}}$, the treated unit's counterfactual predicted outcome is then calculated as $\widehat{Y}_{1t}(0) = \sum_{i=2}^{N+1} \widehat{w}_i Y_{it}(0)$ for $t > T_0$. The weighted average $\widehat{Y}_{1t}(0)$ is often referred to as the *SC* of the treated unit, as it represents how the untreated units can be combined to provide the best counterfactual for the treated unit in the post-treatment period.

When the data contain only information on the outcome of interest, $\mathbf{w}$ is chosen such that the weighted average of the outcomes of the untreated units approximates well the outcome trajectory of the treated unit in the period before the treatment. That is, the weights $\mathbf{w}$ are chosen so that

$$\sum_{i=2}^{N+1} w_i Y_{it}(0) \approx Y_{1t}(0), \qquad \text{for} \quad t = 1, 2, \ldots, T_0,$$

where the meaning of the symbol "$\approx$" varies depending on the specific framework considered. A leading example constrains the weights to be nonnegative and sum to one, and estimates $\mathbf{w}$ by constrained least squares

$$(\widehat{\mathbf{w}}', \widehat{r})' \in \underset{\mathbf{w}\in\mathcal{W}, r\in\mathcal{R}}{\arg\min} \sum_{t=1}^{T_0} (Y_{1t} - Y_{2t}w_2 - \cdots - Y_{(N+1)t}w_{N+1} - r)^2, \tag{2}$$

where r denotes the intercept, and $\mathcal{W}$ and $\mathcal{R}$ denote the corresponding constraint (or feasibility) sets—we give formal definitions in the next subsection.

When the weights are chosen according to Equation (2), the resulting SC will reproduce as closely as possible the outcome trajectory of the treated unit in the pretreatment period. For example, in the Basque terrorism application, this procedure would lead to a synthetic Basque Country that would have a similar per capita GDP to the Basque Country's per capita GDP in the 1955–1975 period when terrorism is not salient.

This outcomes-only version of the SC method, however, cannot guarantee that the resulting SC unit will be similar to the treated unit in any characteristics other than the (pre-treatment) outcome. In some applications, this feature may be undesirable, as researchers may have access to additional characteristics such as baseline covariates or secondary outcomes and may want to also ensure that the SC approximates the treated unit in terms of these additional characteristics. The SC framework can handle this case by including additional equations for these additional characteristics and minimizing the combined loss. In this case, letting $l = 1, 2, \ldots, M$ index the variables that will be "matched" to produce the weights, the minimization problem above can be generalized as follows:

$$(\widehat{\mathbf{w}}', \widehat{\mathbf{r}})' \in \underset{\mathbf{w}\in\mathcal{W}, \mathbf{r}\in\mathcal{R}}{\arg\min} \sum_{l=1}^{M} \sum_{t=1}^{T_0} \upsilon_{t,l} (Y_{1t,l} - Y_{2t,l}w_2 - \cdots - Y_{(N+1)t,l}w_{N+1} - r_l)^2, \tag{3}$$

where $\widehat{\mathbf{r}} = (\widehat{r}_1, \ldots, \widehat{r}_M)'$ and $\{\upsilon_{t,l}\}_{1\le t\le T_0, 1\le l\le M}$ are positive constants reflecting the relative importance of different equations and periods.

For example, in the original Basque terrorism example, Abadie and Gardeazabal (2003) showed that the Basque country differs from the rest of Spain in terms of population density, and they are concerned that pre-terrorism differences in population density may affect economic growth in the post-treatment period. In this case, we can choose the weights $\widehat{\mathbf{w}}$ to ensure not only that the per capita GDP trajectory is similar between the treated unit and the SC unit, but also to ensure that the SC is similar to the treated unit in terms of population density. To implement this multi-equation SC method, we fit equation (3) with two variables ($M = 2$) where $Y_{it,1}$ ($l = 1$) is per capita GDP for region i in year t and $Y_{it,2}$ ($l = 2$) is population density for region i in year t. When $\widehat{\mathbf{w}}$ is chosen this way, the resulting SC will resemble (to the extent that the data allows) the treated unit in terms of both per capita GDP and population density.

Equation (3) can be viewed as a (weighted) combination of M optimization problems in Equation (2), satisfying an additional constraint that the weights $\mathbf{w}$ must be the same across the M equations. For simplicity, we let $\upsilon_{t,l} = 1$ for all t and l, but the analysis below can be applied to the more general case if additional regularity conditions are imposed on $\{\upsilon_{t,l}\}_{1\le t\le T_0, 1\le l\le M}$.

The two cases just discussed (outcomes-only and multi-equation SC frameworks) allow for weakly dependent and cointegrated data, and they also can be generalized further by including covariates in a linear and additive way in Equations (2) or (3). This covariate adjustment would introduce additional parameters to the fit that would not be of primary interest; rather, they would be included to "partial out" the effect of additional covariates.

2.1. General Framework

We now introduce a general framework and further notation that encompass and formalize the two particular examples discussed above as well as other SC approaches in the literature. Our general framework includes the outcomes-only fit and the multi-equation fit (i.e., outcome plus other variables) as particular cases, allowing for covariate adjustment and nonstationary data in a unified way.

Consider SC weights constructed simultaneously for M features of the treated unit, denoted by $\mathbf{A}_l = (a_{1,l}, \cdots, a_{T_0,l})' \in \mathbb{R}^{T_0}$, with index $l = 1, \ldots, M$. For each feature l, there exist $J + K$ variables that can be used to predict or "match" the T_0-dimensional vector $\mathbf{A}_l$. These $J + K$ variables are separated into two groups denoted by $\mathbf{B}_l = (\mathbf{B}_{1,l}, \mathbf{B}_{2,l}, \ldots, \mathbf{B}_{J,l}) \in \mathbb{R}^{T_0\times J}$ and $\mathbf{C}_l = (\mathbf{C}_{1,l}, \ldots, \mathbf{C}_{K,l}) \in \mathbb{R}^{T_0\times K}$, respectively. More precisely, for each j, $\mathbf{B}_{j,l} = (b_{j1,l}, \ldots, b_{jT_0,l})'$ corresponds to the lth feature of the jth unit observed in T_0 pretreatment periods and, for each k, $\mathbf{C}_{k,l} = (c_{k1,l}, \ldots, c_{kT_0,l})'$ is another vector of control variables also possibly used to predict $\mathbf{A}_l$ over the same pre-intervention time span. For ease of notation, we let $d = J + KM$.

The goal of the SC method is to search for a vector of common weights $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^J$ across the M features and a vector of coefficients $\mathbf{r} \in \mathcal{R} \subseteq \mathbb{R}^{KM}$, such that the linear

combination of $\mathbf{B}_l$ and $\mathbf{C}_l$ "matches" $\mathbf{A}_l$ as close as possible, for all $1 \leq l \leq M$. This goal is typically achieved via the following optimization problem:

$$\widehat{\boldsymbol{\beta}} := (\widehat{\mathbf{w}}', \widehat{\mathbf{r}}')' \in \underset{\mathbf{w}\in\mathcal{W},\, \mathbf{r}\in\mathcal{R}}{\arg\min}\ (\mathbf{A} - \mathbf{B}\mathbf{w} - \mathbf{C}\mathbf{r})'(\mathbf{A} - \mathbf{B}\mathbf{w} - \mathbf{C}\mathbf{r}) \quad (4)$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_M \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_M \end{bmatrix},$$

and where the feasibility sets $\mathcal{W}$ and $\mathcal{R}$ capture the restrictions imposed. (For simplicity we do not introduce an explicit reweighting of the M equations, but recall that this extension is possible.) This framework encompasses multiple prior SC formalizations in the literature, which differ in whether they include additional covariates, whether the data is assumed to be stationary, and the particular choice of constraint sets $\mathcal{W}$ and $\mathcal{R}$ used, among other possibilities.

The following list provides some examples of different constraint sets used in practice, where $||\cdot||_p$ denotes the L_p vector norm and Q and α are tuning parameters.

- Abadie, Diamond, and Hainmueller (2010): $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N_+ : ||\mathbf{w}||_1 = 1\}$ and $\mathcal{R} = \{0\}$.
- Hsiao, Steve Ching and Ki Wan (2012): $\mathcal{W} = \mathbb{R}^N$ and $\mathcal{R} = \mathbb{R}$.
- Ferman and Pinto (2021): $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N_+ : ||\mathbf{w}||_1 = 1\}$ and $\mathcal{R} = \mathbb{R}$.
- Chernozhukov, Wüthrich, and Zhu (2021b): $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N : ||\mathbf{w}||_1 \leq 1\}$ and $\mathcal{R} = \mathbb{R}$.
- Amjad, Shah, and Shen (2018): $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N : ||\mathbf{w}||_2 \leq Q\}$ and $\mathcal{R} = \{0\}$.
- Arkhangelsky et al. (2021): $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N : ||\mathbf{w}||_2 \leq Q, ||\mathbf{w}||_1 = 1\}$ and $\mathcal{R} = \mathbb{R}$.
- Doudchenko and Imbens (2016): $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N : \frac{1-\alpha}{2}||\mathbf{w}||_2^2 + \alpha||\mathbf{w}||_1 \leq Q\}$ and $\mathcal{R} = \mathbb{R}$.

In some applications the intercept in Equation (4) is removed by demeaning the data before the analysis. Section 4.1 discusses in detail the outcomes-only case, as well as the multi-equation case where the researcher "matches" on pretreatment characteristics and pre-intervention outcomes simultaneously. That section also deals with stationary weakly dependant data, and nonstationary data (i.e., cointegration system).

For example, the outcomes-only setup can be obtained as a particular case of Equation (4) with $M = 1$ (there is only one feature to match on), $J = N$ (there are N units in the donor pool), and $K = 1$ (there is an intercept). Then, $\mathbf{A}_1 = (Y_{11}, Y_{12}, \ldots, Y_{1T_0})'$, $\mathbf{B}_{j,1} = (Y_{(j+1)1}, Y_{(j+1)2}, \ldots, Y_{(j+1)T_0})'$, $\mathbf{C}_{j,1} = (1, 1, \ldots, 1)'$, and Equation (4) reduces to the (possibly constrained) optimization problem (2). The multi-equation setup with one outcome and one covariate can be obtained similarly by setting $M = 2$ (there are two features to match on), $J = N$ (N units in the donor pool), and $K = 1$ (there is an intercept), which reduces to Equation (3).

To further understand our proposed inference approach, we define the pseudo-true values $\mathbf{w}_0$ and $\mathbf{r}_0$ relative to a sigma field $\mathscr{H}$:

$$\boldsymbol{\beta}_0 := (\mathbf{w}_0', \mathbf{r}_0')' = \underset{\mathbf{w}\in\mathcal{W},\, \mathbf{r}\in\mathcal{R}}{\arg\min}\ \mathbb{E}[(\mathbf{A}-\mathbf{B}\mathbf{w}-\mathbf{C}\mathbf{r})'(\mathbf{A}-\mathbf{B}\mathbf{w}-\mathbf{C}\mathbf{r})|\mathscr{H}], \quad (5)$$

and thus write

$$\mathbf{A} = \mathbf{B}\mathbf{w}_0 + \mathbf{C}\mathbf{r}_0 + \mathbf{U}, \qquad \mathbf{w}_0 \in \mathcal{W}, \qquad \mathbf{r}_0 \in \mathcal{R}, \quad (6)$$

where $\mathbf{U} = (u_{1,1}, \ldots, u_{T_0,1}, \ldots, u_{1,M}, \cdots, u_{T_0,M})' \in \mathbb{R}^{T_0 M}$ is the corresponding pseudo-true residual relative to a sigma field $\mathscr{H}$. That is, $\mathbf{w}_0$ and $\mathbf{r}_0$ are the mean square error estimands associated with the (possibly constrained) best linear prediction coefficients $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{r}}$ conditional on $\mathscr{H}$. Importantly, we *do not* attach any structural meaning to Equation (6). The population vectors $\mathbf{w}_0$ and $\mathbf{r}_0$ are (conditional) pseudo-true values whose meaning should be understood in context, and are determined by the assumptions imposed on the data generating process. In particular, with strong parametric functional form assumptions or rich enough nonparametric basis expansions, Equation (6) may be viewed as a representation (or approximation) of $\mathbb{E}[\mathbf{A}|\mathbf{B}, \mathbf{C}, \mathscr{H}]$. In such cases, $\mathbb{E}[\mathbf{U}|\mathbf{B}, \mathbf{C}, \mathscr{H}] = \mathbf{0}$ or, at least, $\mathbb{E}[\mathbf{U}|\mathbf{B}, \mathbf{C}, \mathscr{H}]$ is taken to be "small". Alternatively, if the (population, conditional) linear projection coefficients lie on $\mathcal{W} \times \mathcal{R}$, that is, the constraints imposed by $\mathcal{W}$ and $\mathcal{R}$ in Equation (5) are not binding, then Equation (6) represents the best linear approximation of $\mathbf{A}$ based on $(\mathbf{B}, \mathbf{C})$, conditional on $\mathscr{H}$. In this scenario, $\mathbf{U}$ is uncorrelated with $(\mathbf{B}, \mathbf{C})$, conditional on $\mathscr{H}$. Most importantly, in general, $\mathbf{U}$ may not be mean zero due to the (binding) constraints imposed in the construction of $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{r}}$.

Given estimated weights $\widehat{\mathbf{w}}$ and coefficients $\widehat{\mathbf{r}}$, the post-treatment counterfactual outcome for the treated unit is predicted by

$$\widehat{Y}_{1T}(0) = \mathbf{x}_T'\widehat{\mathbf{w}} + \mathbf{g}_T'\widehat{\mathbf{r}} = \mathbf{p}_T'\widehat{\boldsymbol{\beta}}, \quad \mathbf{p}_T := (\mathbf{x}_T', \mathbf{g}_T')', \quad T > T_0,$$

where $\mathbf{x}_T \in \mathbb{R}^N$ is a vector of predictors for control units observed in time T and $\mathbf{g}_T \in \mathbb{R}^{KM}$ is another set of user-specified predictors observed at time T. Variables included in $\mathbf{x}_T$ and $\mathbf{g}_T$ need not be the same as those in $\mathbf{B}$ and $\mathbf{C}$, but will be part of the sigma field $\mathscr{H}$, as explained in more detail in the next section. Therefore, from this perspective, our focus is on conditional inference. We decompose the potential outcome of the treated unit accordingly:

$$Y_{1T}(0) \equiv \mathbf{x}_T'\mathbf{w}_0 + \mathbf{g}_T'\mathbf{r}_0 + e_T = \mathbf{p}_T'\boldsymbol{\beta}_0 + e_T, \qquad T > T_0, \quad (7)$$

where e_T is defined by construction. In our analysis, $\mathbf{w}_0$ and $\mathbf{r}_0$ are assumed to be possibly random elements around which $\widehat{\mathbf{w}}$ and $\widehat{\mathbf{r}}$ are concentrating in probability, respectively, which is why we called them pseudo-true values.

The distance between the estimated treatment effect on the treated and the target population one is

$$\begin{aligned} \widehat{\tau}_T - \tau_T &= \Big(Y_{1T}(1) - \widehat{Y}_{1T}(0)\Big) - \Big(Y_{1T}(1) - Y_{1T}(0)\Big) \\ &= Y_{1T}(0) - \widehat{Y}_{1T}(0). \end{aligned} \quad (8)$$

Within the SC framework, we view the quantity of interest τ_T as a random variable, and hence we refrain from calling it a "parameter." Consequently, we call $\widehat{\tau}_T$ a prediction of τ_T rather than an "estimator" of it, and focus on building prediction intervals rather than confidence intervals.

3. Prediction Intervals

Given the generic framework introduced in the previous section, we now present our proposed prediction intervals for τ_T. See Vovk (2012), Chernozhukov, Wüthrich, and Zhu (2021a); Chernozhukov, Wüthrich, and Zhu (2021b), and references therein, for recent papers on (conditional) prediction intervals and related methods. Let $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ be random quantities defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, and $\mathscr{H} \subseteq \mathscr{F}$ be a sub-σ-field. For some $\alpha, \pi \in (0,1)$, we say a random interval $\mathcal{I}$ is an (α,π)-valid $\mathscr{H}$-conditional prediction interval for τ_T if

$$\mathbb{P}\Big\{\mathbb{P}\big[\tau_T \in \mathcal{I} \,\big|\, \mathscr{H}\big] \geq 1-\alpha\Big\} \geq 1-\pi. \tag{9}$$

If $\mathscr{H}$ is the trivial σ-field over Ω, then $\mathcal{I}$ reduces to an unconditional prediction interval for τ_T. In the general case, $\mathcal{I}$ is an $\mathscr{H}$-conditionally (α,π)-valid prediction interval: the conditional coverage probability of $\mathcal{I}$ is at least $(1-\alpha)$, which holds with probability over $\mathscr{H}$ at least $(1-\pi)$. In practice, $(1-\alpha)$ is a desired confidence level chosen by users, say 95%, and π is a "small" number that depends on the sample size and typically goes to zero in some asymptotic sense. In this paper, all the results are valid for all T_0 large enough, with the associated probability loss π characterized precisely. Thus, we say that the conditional coverage of the prediction interval $\mathcal{I}$ is at least $(1-\alpha)$ with high probability, or that the conditional prediction interval offers finite-sample probability guarantees. Our results imply $\pi \to 0$ as $T_0 \to \infty$, but no limits or asymptotic arguments are used in this paper.

An asymptotic analogue to the above definition (9) would be $\mathbb{P}(\tau_T \in \mathcal{I}|\mathscr{H}) \geq 1-\alpha-o_{\mathbb{P}}(1)$ or, perhaps, $\mathbb{P}(\tau_T \in \mathcal{I}|\mathscr{H}) \to_{\mathbb{P}} 1-\alpha$, where the probability limit is taken as the sample size grows to infinity (e.g., as $T_0 \to \infty$). In this case, we say $\mathcal{I}$ is an $\mathscr{H}$-conditional prediction interval for τ_T that is asymptotically valid with coverage probability (at least) $(1-\alpha)$. This is a weaker property because it does not offer any finite-sample probability guarantees for the (conditional) coverage of the prediction interval.

We employ the following lemma to construct valid, conditional prediction intervals in the sense of Equation (9). This lemma follows from the union bound applied to $\widehat{\tau}_T - \tau_T = \mathbf{p}_T'(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}) + e_T$.

Lemma 1 (Prediction Interval). Suppose that there exist $M_{1,\mathrm{L}}$, $M_{1,\mathrm{U}}$, $M_{2,\mathrm{L}}$ and $M_{2,\mathrm{U}}$, possibly depending on $\alpha_1, \alpha_2, \pi_1, \pi_2 \in (0,1)$ and the conditioning σ-field $\mathscr{H}$, such that

$$\mathbb{P}\Big\{\mathbb{P}\big[M_{1,\mathrm{L}} \leq \mathbf{p}_T'(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}) \leq M_{1,\mathrm{U}} \,\big|\, \mathscr{H}\big] \geq 1-\alpha_1\Big\} \geq 1-\pi_1, \quad \text{and}$$
$$\mathbb{P}\Big\{\mathbb{P}\big[M_{2,\mathrm{L}} \leq \quad e_T \quad \leq M_{2,\mathrm{U}} \,\big|\, \mathscr{H}\big] \geq 1-\alpha_2\Big\} \geq 1-\pi_2.$$

Then, $\mathbb{P}\Big\{\mathbb{P}\big[\widehat{\tau}_T - M_{1,\mathrm{U}} - M_{2,\mathrm{U}} \leq \tau_T \leq \widehat{\tau}_T - M_{1,\mathrm{L}} - M_{2,\mathrm{L}} \big| \mathscr{H}\big] \geq 1-\alpha_1-\alpha_2\Big\} \geq 1-\pi_1-\pi_2$.

This lemma provides a simple way to construct an $\mathscr{H}$-conditional prediction interval enjoying (α,π)-validity with $\alpha = \alpha_1 + \alpha_2$ and $\pi = \pi_1 + \pi_2$:

$$\mathcal{I} = \Big[\widehat{\tau}_T - M_{1,\mathrm{U}} - M_{2,\mathrm{U}}\,,\ \widehat{\tau}_T - M_{1,\mathrm{L}} - M_{2,\mathrm{L}}\Big],$$

for appropriate choices of $M_{1,\mathrm{L}}$, $M_{1,\mathrm{U}}$, $M_{2,\mathrm{L}}$ and $M_{2,\mathrm{U}}$ and conditioning sigma field. In this paper, we consider conditional prediction intervals with $\mathscr{H} = \sigma(\mathbf{B}, \mathbf{C}, \mathbf{x}_T, \mathbf{g}_T)$ and focus on building a probability bound for each of the two terms, $\mathbf{p}_T'(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})$ and e_T, separately, and then combine them to build an overall probability bound via Lemma 1. In the decomposition leading to the prediction interval construction, we interpret $\mathbf{p}_T'(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})$ as capturing the in-sample uncertainty coming from constructing the SC weights using pretreatment information, and e_T the out-of-sample uncertainty coming from misspecification along with any additional noise occurring at the post-treatment period $T > T_0$. The next two subsections are devoted to handle each of these terms, respectively.

Remark 1 (Prediction Interval for $Y_{1T}(0)$). Once the $\mathscr{H}$-conditionally (α,π)-valid prediction interval $\mathcal{I}$ for τ_T is constructed, an analogous prediction interval for the counterfactual outcome of the treated unit in the post-treatment period T, $Y_{1T}(0)$, is also readily available. To be precise, using Equation (1), it follows that

$$\mathbb{P}\Big\{\mathbb{P}\big[Y_{1T}(1) - Y_{1T}(0) \in \mathcal{I} \big| \mathscr{H}\big] \geq 1-\alpha\Big\} \geq 1-\pi,$$

that is, $\big[M_{1,\mathrm{L}} + M_{2,\mathrm{L}} + \widehat{Y}_{1T}(0)\,,\ M_{1,\mathrm{U}} + M_{2,\mathrm{U}} + \widehat{Y}_{1T}(0)\big]$ is a conditionally valid prediction interval for $Y_{1T}(0)$.

4. In-Sample Uncertainty

We first quantify the in-sample uncertainty coming from $\mathbf{p}_T'(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})$, thereby providing methods to determine $(M_{1,\mathrm{L}}, M_{1,\mathrm{U}})$ and their probability guarantees (α_1, π_1) in Lemma 1. Let $\mathbf{Z} = (\mathbf{B}, \mathbf{C})$, $\mathbf{D}$ be a nonnegative diagonal (scaling) matrix of size d, possibly depending on the pretreatment sample size T_0, and recall that $\mathscr{H} = \sigma(\mathbf{B}, \mathbf{C}, \mathbf{x}_T, \mathbf{g}_T)$. Because $\widehat{\boldsymbol{\beta}}$ solves (4), we can define $\widehat{\boldsymbol{\delta}} := \mathbf{D}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ as the optimizer of the centered criterion function:

$$\widehat{\boldsymbol{\delta}} = \operatorname*{arg\,min}_{\boldsymbol{\delta} \in \Delta} \big\{\boldsymbol{\delta}'\widehat{\mathbf{Q}}\boldsymbol{\delta} - 2\widehat{\boldsymbol{\gamma}}'\boldsymbol{\delta}\big\},$$

where $\widehat{\mathbf{Q}} = \mathbf{D}^{-1}\mathbf{Z}'\mathbf{Z}\mathbf{D}^{-1}$, $\widehat{\boldsymbol{\gamma}}' = \mathbf{U}'\mathbf{Z}\mathbf{D}^{-1}$, and $\Delta = \{\boldsymbol{h} \in \mathbb{R}^d : \boldsymbol{h} = \mathbf{D}(\boldsymbol{\beta} - \boldsymbol{\beta}_0),\ \boldsymbol{\beta} \in \mathcal{W} \times \mathcal{R}\}$.

The following lemma, which holds whether or not $\boldsymbol{\gamma} := \mathbb{E}[\widehat{\boldsymbol{\gamma}}|\mathscr{H}] = \mathbf{0}$, is a key building block for our prediction interval construction.

Lemma 2 (Optimization Bounds). Fix $\widehat{\mathbf{Q}}$ and $\mathbf{p}_T$. Assume $\mathcal{W}$ and $\mathcal{R}$ are convex, and let $\widehat{\boldsymbol{\beta}}$ in Equation (4) and $\boldsymbol{\beta}_0$ in Equation (5) exist. Then,

$$\varsigma_{\mathrm{L}} := \inf_{\boldsymbol{\delta} \in \mathcal{M}_{\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}}} \mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} \ \leq\ \mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}}$$
$$\leq \sup_{\boldsymbol{\delta} \in \mathcal{M}_{\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}}} \mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} =: \varsigma_{\mathrm{U}},$$

where $\mathcal{M}_{\boldsymbol{\xi}} = \{\boldsymbol{\delta} \in \Delta : \boldsymbol{\delta}'\widehat{\mathbf{Q}}\boldsymbol{\delta} - 2\boldsymbol{\xi}'\boldsymbol{\delta} \leq 0\}$. Furthermore, for any $\kappa \in \mathbb{R}$,

$$\Big\{\boldsymbol{\xi} \in \mathbb{R}^d : \inf_{\boldsymbol{\delta} \in \mathcal{M}_{\boldsymbol{\xi}}} \mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} \geq \kappa\Big\} \quad \text{and}$$
$$\Big\{\boldsymbol{\xi} \in \mathbb{R}^d : \sup_{\boldsymbol{\delta} \in \mathcal{M}_{\boldsymbol{\xi}}} \mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} \leq \kappa\Big\}$$

are convex sets.

This lemma does not involve probabilistic statements, but rather follows from basic features of constrained least-square optimization. In particular, simple bounds on $\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}}$ can be deduced based on the basic inequality from optimization $\widehat{\boldsymbol{\delta}}'\widehat{\mathbf{Q}}\widehat{\boldsymbol{\delta}} - 2(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})'\widehat{\boldsymbol{\delta}} \leq 0$, and the fact that any solution must satisfy the constraints imposed in Equation (4), that is, $\widehat{\boldsymbol{\delta}} \in \Delta$. The second part of the lemma establishes that the set of possible localization values ($\boldsymbol{\xi}$) determining the feasibility set ($\mathcal{M}_{\boldsymbol{\xi}}$) of the bounding (random) quantities ς_{L} and ς_{U} in Lemma 2 (when $\boldsymbol{\xi} = \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$) form (random) convex sets.

Conditional on $\mathscr{H}$, the set $\mathcal{M}_{\boldsymbol{\xi}}$ is not random due to $\widehat{\mathbf{Q}}$, which is random only unconditionally. As a consequence, conditional on $\mathscr{H}$, both $\mathcal{M}_{\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}}$ and $\{\mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} : \boldsymbol{\delta} \in \mathcal{M}_{\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}}\}$ are random sets only because $\widehat{\boldsymbol{\gamma}}$ is a random quantity and, accordingly, ς_{L} and ς_{U} are random variables defined by a random set. If the conditional distributions of ς_{L} and ς_{U} were known, we could take their quantiles as lower and upper bounds for the quantiles of the conditional distribution of $\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}} = \mathbf{p}_T'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, thereby transforming the first conclusion of Lemma 2 into a probabilistic statement. However, this approach requires knowledge of the conditional (on $\mathscr{H}$) distribution of the bounding random variables ς_{L} and ς_{U}. The convexity properties also established in Lemma 2 allow us to provide precise bounds on the desired conditional distribution of the bounding random variables using Berry-Esseen bounds for convex sets (Raič 2019).

The following theorem formalizes our first main result based on Lemma 2. We only present the result for the upper bound to conserve space, but the analogous result holds for the lower bound. See Remarks SA-2.1–SA-2.3 in the supplemental appendix for more details. Let $\Sigma = \mathbb{V}[\widehat{\boldsymbol{\gamma}}|\mathscr{H}]$ and $\Sigma^{-1/2}\mathbf{D}^{-1}\mathbf{Z}' = (\tilde{\mathbf{z}}_{1,1},\cdots,\tilde{\mathbf{z}}_{T_0,1},\ldots,\tilde{\mathbf{z}}_{1,M},\ldots,\tilde{\mathbf{z}}_{T_0,M})$. In addition, let $||\cdot||$ denote the spectral matrix norm (so that $||\cdot|| = ||\cdot||_2$ for vectors).

Theorem 1 (Distributional Approximation, Independent Case). Assume $\mathcal{W}$ and $\mathcal{R}$ are convex, $\widehat{\boldsymbol{\beta}}$ in Equation (4) and $\boldsymbol{\beta}_0$ in Equation (5) exist, and $\mathscr{H} = \sigma(\mathbf{B},\mathbf{C},\mathbf{x}_T,\mathbf{g}_T)$. In addition, for some finite nonnegative constants ϵ_γ and π_γ, the following conditions hold:

(T1.i) $\mathbf{u}_t = (u_{t,1},\ldots,u_{t,M})'$ is independent over t conditional on $\mathscr{H}$;
(T1.ii) $\mathbb{P}\{\sum_{t=1}^{T_0}\mathbb{E}[||\sum_{l=1}^{M}\tilde{\mathbf{z}}_{t,l}(u_{t,l} - \mathbb{E}[u_{t,l}|\mathscr{H}])||^3|\mathscr{H}] \leq \epsilon_\gamma(42(d^{1/4}+16))^{-1}\} \geq 1-\pi_\gamma$.

Then,

$$\mathbb{P}\Big[\mathbb{P}\big(\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}} \leq \mathfrak{c}^\dagger(1-\alpha)\big|\mathscr{H}\big) \geq 1-(\alpha+\epsilon_\gamma)\Big] \geq 1-\pi_\gamma,$$

where $\mathfrak{c}^\dagger(1-\alpha)$ denotes the $(1-\alpha)$-quantile of $\varsigma_{\mathrm{U}}^\dagger = \sup\{\mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} : \boldsymbol{\delta} \in \mathcal{M}_{\mathbf{G}}\}$ conditional on $\mathscr{H}$, with $\mathcal{M}_{\mathbf{G}} = \{\boldsymbol{\delta} \in \Delta : \ell^\dagger(\boldsymbol{\delta}) \leq 0\}$, $\ell^\dagger(\boldsymbol{\delta}) := \boldsymbol{\delta}'\widehat{\mathbf{Q}}\boldsymbol{\delta} - 2\mathbf{G}'\boldsymbol{\delta}$, and $\mathbf{G}|\mathscr{H} \sim \mathsf{N}(\mathbf{0},\Sigma)$.

This theorem is established under two high-level conditions. Condition (T1.i) imposes independence across time for the pseudo-residuals underlying the population analogue construction of the SC weights in (6). In Appendix A we relax this requirement by allowing for weak dependence across time via a β-mixing condition (Doukhan 2012), but to avoid untidy conditions we focus on the independent case here. Importantly, even in this case, Theorem 1 covers nonstationarity in the outcome variable (via a cointegration relationship). See Section 4.1 for different examples with independent, weakly stationary, and nonstationary data.

The second high-level requirement in Theorem 1 helps control the (distributional) distance between $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ and the Gaussian random vector $\mathsf{N}(\mathbf{0},\Sigma)$, conditionally on $\mathscr{H}$, as well as the unconditional probability loss π_γ. Condition (T1.ii) can be verified in a variety of ways depending on the dependence structure imposed on the data and other regularity conditions, as we illustrate in Section 4.1. For instance, two sufficient conditions are: $\max_{1\leq l\leq M}\max_{1\leq t\leq T_0}\mathbb{E}[|u_{t,l} - \mathbb{E}[u_{t,l}|\mathscr{H}]|^3|\mathscr{H}] \leq \eta$, a.s. on $\mathscr{H}$ for some constant $\eta > 0$, and $\mathbb{P}\{\sum_{t=1}^{T_0}\sum_{l=1}^{M}||\tilde{\mathbf{z}}_{t,l}||^3 \leq \epsilon_\gamma(42(d^{1/4}+16)\eta M^2)^{-1}\} \geq 1-\pi_\gamma$.

Since $\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}} = \mathbf{p}_T'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, Theorem 1 could immediately be applied to construct valid $M_{1,\mathrm{L}}$ and $M_{1,\mathrm{U}}$ in Lemma 1 if $\Sigma = \mathbb{V}[\widehat{\boldsymbol{\gamma}}|\mathscr{H}]$ was known. Thus, to finalize the in-sample uncertainty quantification we discuss a feasible simulation-based approximation for the critical value $\mathfrak{c}^\dagger(1-\alpha)$. To describe such approach, define a simulation-based criterion function conditional on the data

$$\ell^\star(\boldsymbol{\delta}) = \boldsymbol{\delta}'\widehat{\mathbf{Q}}\boldsymbol{\delta} - 2(\mathbf{G}^\star)'\boldsymbol{\delta}, \qquad \mathbf{G}^\star \sim \mathsf{N}(0,\widehat{\Sigma}),$$

where $\widehat{\Sigma}$ is some estimate of Σ. The form of $\widehat{\Sigma}$ depends on the specific dependence structure underlying the data and related regularity conditions, as we illustrate in Section 4.1. Naturally, the important high-level requirement is that $\widehat{\Sigma}$ should concentrate around Σ with known probability; see Theorem 2 for the precise statement. In addition, the constraint set used in the simulation has to be properly defined to account for the parameters being possibly near or at the boundary, so that it mimics the local geometry of Δ. Specifically, let $\Delta^\star$ denote the constraint set used in simulation. We require that

$$\Delta^\star \cap \mathcal{B}(\mathbf{0},\varepsilon) = \Delta \cap \mathcal{B}(\mathbf{0},\varepsilon), \qquad \text{for some } \varepsilon > 0, \tag{10}$$

where $\mathcal{B}(\mathbf{0},\varepsilon)$ is an ε-neighborhood around zero. We say $\Delta^\star$ is *locally equal to* Δ if Equation (10) is satisfied. Consequently, searching for the desired region under constraints in $\Delta^\star$ is almost equivalent to doing so under constraints in Δ. We discuss below more implementation details.

The next theorem establishes the validity of our proposed simulation-based inference method and provides the associated probability guarantees, under high-level conditions. Let $||\cdot||_{\mathrm{F}}$ denote the Frobenius matrix norm (so that $||\cdot||_{\mathrm{F}} = ||\cdot|| = ||\cdot||_2$ for vectors), and $\mathbf{I}_q$ the identity matrix of size q for an integer $q > 0$.

Theorem 2 (Plug-in Approximation). Assume $\mathcal{W}$ and $\mathcal{R}$ are convex, $\widehat{\boldsymbol{\beta}}$ in Equation (4) and $\boldsymbol{\beta}_0$ in Equation (5) exist, and $\mathscr{H} = \sigma(\mathbf{B},\mathbf{C},\mathbf{x}_T,\mathbf{g}_T)$. In addition, for some finite nonnegative constants ϵ_γ, π_γ, $\varpi_\delta^\star$, $\epsilon_\delta^\star$, $\pi_\delta^\star$, $\epsilon_\Delta^\star$, $\pi_\Delta^\star$, $\epsilon_{\gamma,1}^\star$, $\epsilon_{\gamma,2}^\star$ and $\pi_\gamma^\star$, the following conditions hold:

(T2.i) $\mathbb{P}[\mathbb{P}(\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}} \leq \mathfrak{c}^\dagger(1-\alpha)|\mathscr{H}) \geq 1-\alpha-\epsilon_\gamma] \geq 1-\pi_\gamma$;
(T2.ii) $\mathbb{P}[\mathbb{P}(\sup\{||\boldsymbol{\delta}|| : \boldsymbol{\delta} \in \mathcal{M}_{\mathbf{G}}\} \leq \varpi_\delta^\star|\mathscr{H}) \geq 1-\epsilon_\delta^\star] \geq 1-\pi_\delta^\star$;
(T2.iii) $\mathbb{P}[\mathbb{P}(\Delta^\star \text{ is locally equal to } \Delta\ |\mathscr{H}) \geq 1-\epsilon_\Delta^\star] \geq 1-\pi_\Delta^\star$ for $\varepsilon = \varpi_\delta^\star$ in (10);

(T2.iv) $\mathbb{P}[\mathbb{P}(||\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}-\mathbf{I}_d||_{\mathbb{F}} \leq 2\epsilon^\star_{\gamma,1}|\mathscr{H}) \geq 1-\epsilon^\star_{\gamma,2}] \geq 1-\pi^\star_\gamma$.

Then, for $\epsilon^\star_{\gamma,1} \in [0, 1/4]$,

$$\mathbb{P}\Big[\mathbb{P}\big(\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}} \leq \mathfrak{c}^\star(1-\alpha)\big|\mathscr{H}\big) \geq 1-\alpha-\epsilon\Big] \geq 1-\pi,$$

where $\epsilon = \epsilon_\gamma+\epsilon^\star_{\gamma,1}+\epsilon^\star_{\gamma,2}+\epsilon^\star_\delta+\epsilon^\star_\Delta, \pi = \pi_\gamma+\pi^\star_\gamma+\pi^\star_\delta+\pi^\star_\Delta$, and $\mathfrak{c}^\star(1-\alpha)$ denotes the $(1-\alpha)$-quantile of $\varsigma^\star_{\mathrm{U}} := \sup\{\mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} : \boldsymbol{\delta} \in \Delta^\star, \ell^\star(\boldsymbol{\delta}) \leq 0\}$, conditional on the data.

This theorem gives a feasible, simulation-based approach to determine valid $M_{1,\mathrm{L}}$ and $M_{1,\mathrm{U}}$ in Lemma 1, with precise coverage probability guarantees. The first high-level Condition (T2.i) in Theorem 2 takes as a starting point the conclusion of Theorem 1 or, alternatively, the conclusion of Theorem A in the appendix when the data are assumed to exhibit weak dependence via a β-mixing condition. The other three high-level conditions in Theorem 2 are intuitive. Conditions (T2.ii) and (T2.iii) control the local geometry of the simulation feasibility set, as discussed earlier, while Condition (T2.iv) requires $\widehat{\Sigma}$ to be a "good" approximation of Σ, in the sense that $\widehat{\Sigma}$ concentrates in probability around Σ with well-controlled errors. Importantly, Theorem 2 is carefully crafted to accommodate both Theorem 1 (independent data) and Theorem A in the appendix (weakly dependent time series data) in a unified way. The next section illustrates different cases with practically relevant examples, and gives precise primitive conditions.

4.1. Examples

We consider the standard SC constraints $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^N_+ : ||\mathbf{w}||_1 = 1\}$ and $\mathcal{R} = \mathbb{R}^{KM}$. For simulation-based inference, we define explicitly a relaxed constraint set based on the original estimated coefficients $\widehat{\boldsymbol{\beta}}$: $\Delta^\star = \{\mathbf{D}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}^\star) : \boldsymbol{\beta} = (\mathbf{w}',\mathbf{r}')', \mathbf{w} \in \mathbb{R}^N_+, ||\mathbf{w}||_1 = ||\widehat{\mathbf{w}}^\star||_1\}$, where $\widehat{\boldsymbol{\beta}}^\star = (\widehat{\mathbf{w}}^{\star\prime},\widehat{\mathbf{r}}')'$, $\widehat{\mathbf{w}}^\star = (\widehat{\omega}^\star_2,\ldots,\widehat{\omega}^\star_{N+1})'$, $\widehat{\omega}^\star_j = \widehat{\omega}_j 1(|\widehat{\omega}_j| > \varrho)$, and ϱ is a tuning parameter that ensures the constraint set in the simulation world preserves the local geometry of Δ. Moreover, we set $\mathbf{x}_T = (Y_{2T}(0),\ldots,Y_{(N+1)T}(0))'$ as it is common in the SC literature. Other SC methods that vary these choices, including the other constraint sets $\mathcal{W}$ discussed previously, can be handled analogously, but we do not discuss them in this paper due to space limitations. Finally, in the remaining of this article, we let $\mathfrak{C}$, $\mathfrak{C}^\star$ and $\mathfrak{c}$, with various sub-indexes, denote nonnegative finite constants not depending on T_0. In simple cases, we give the exact expression of these constants, while in other cases they can be characterized from the proofs of the results. Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ be the minimum and the maximum eigenvalues of a generic square matrix $\mathbf{M}$.

4.1.1. Outcomes-only

We start with the simplest possible example already introduced in Section 2. The SC weights are constructed based on past outcomes only, and the model allows for an intercept. Thus, the working model simplifies to

$$a_t = \mathbf{b}_t'\mathbf{w}_0 + r_0 + u_t, \qquad t = 1,\ldots,T_0,$$

where $a_t := Y_{1t}(0)$, $\mathbf{b}_t := (Y_{2t}(0), Y_{3t}(0),\ldots,Y_{(N+1)t}(0))'$, and with $M = 1$, $K = 1$, and $d = N+1$. Recall that $\mathbf{w}_0 = (w_{0,1}, w_{0,2},\ldots,w_{0,J})'$ is defined in Equation (5), and let $\mathbf{z}_t = (\mathbf{b}_t', 1)'$, $\boldsymbol{\beta}_0 = (\mathbf{w}_0', r_0)'$. We further assume independent sampling across time, and thus set $\mathbf{D} = T_0^{1/2}\mathbf{I}_d$. A natural variance estimator is

$$\widehat{\Sigma} = \frac{1}{T_0}\sum_{t=1}^{T_0}\mathbf{z}_t\mathbf{z}_t'(\widehat{u}_t - \widehat{\mathbb{E}}[u_t|\mathbf{b}_t])^2,$$

where $\widehat{u}_t = a_t - \mathbf{z}_t'\widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbb{E}}[u_t|\mathbf{b}_t]$ denotes some estimate of the conditional mean of the pseudo-residuals.

Theorem SA-1 in the supplemental appendix gives precise primitive conditions to verify the high-level conditions of Theorems 1 and 2. In particular, assuming that $\{\mathbf{z}_t, u_t\}_{t=1}^T$ is iid over $t = 1,\ldots,T_0$, and that $\max_{1\leq t\leq T_0}\mathbb{E}[|u_t|^3|\mathbf{B}] \leq \bar{\eta}_1$ a.s. on $\sigma(\mathbf{B})$ and $\mathbb{E}[||\mathbf{z}_t||^6] \leq \bar{\eta}_2$, $\min_{1\leq t\leq T_0}\mathbb{V}[u_t|\mathbf{B}] \geq \underline{\eta}_1$ a.s. on $\sigma(\mathbf{B})$, and $\lambda_{\min}(\mathbb{E}[\mathbf{z}_t\mathbf{z}_t']) \geq \underline{\eta}_2$, for finite nonnegative constants $\bar{\eta}_1$, $\bar{\eta}_2$, $\underline{\eta}_1$ and $\underline{\eta}_2$, we show that the conditions of Theorem 1 hold with $\pi_\gamma = \mathfrak{C}_\pi T_0^{-1}$ and $\epsilon_\gamma = \mathfrak{C}_\epsilon T_0^{-1/2}$, where $\mathfrak{C}_\pi = \frac{d}{\bar{\eta}_2} + \frac{4d^4\bar{\eta}_2}{\underline{\eta}_2^2}$ and $\mathfrak{C}_\epsilon = 42(d^{1/4}+16)\frac{2^{5/2}d^{3/2}\bar{\eta}_1\bar{\eta}_2}{(\underline{\eta}_1\underline{\eta}_2)^{3/2}}$. Furthermore, under additional primitive conditions, we also show that the conditions of Theorem 2 hold with precise nonasymptotic probability bounds characterized in the proof.

Theorem SA-1 characterizes precisely the probability guarantees for the in-sample prediction—that is, the precise values of α_1 and π_1 in Lemma 1 obtained via Theorems 1 and 2. The conditions imposed are primitive (e.g., moment bounds and rank conditions), with perhaps the exception of conditions (SA-1.iii) and (SA-1.v) in Theorem SA-1 in the supplemental appendix. Specifically, Condition (SA-1.iii) requires $\varrho = \varpi^\star_\delta/\sqrt{T_0}$ and $\mathbb{P}(\min\{|w_{0,j}| : w_{0,j} \neq 0\} \geq \varrho) \geq 1-\pi^\star_w$, for nonnegative constants $\varpi^\star_\delta$ and $\pi^\star_w$, which is also primitive insofar it relates to the separation from zero of the nonzero (possibly random) coefficients $\mathbf{w}_0$ entering the best linear approximation (5), which is a standard (sparsity-type) assumption in the literature of constrained least-square estimation. On the other hand, Condition (SA-1.v) requires $\mathbb{P}[\mathbb{P}(\max_{1\leq t\leq T_0}|\widehat{\mathbb{E}}[u_t|\mathbf{b}_t] - \mathbb{E}[u_t|\mathbf{b}_t]| \leq \varpi^\star_u|\mathscr{H}) \geq 1-\epsilon^\star_u] \geq 1-\pi^\star_u$, for nonnegative constants $\varpi^\star_u$, $\epsilon^\star_u$ and $\pi^\star_u$, which is purposely not as primitive (but still easily interpretable) because it is meant to cover many different approximation approaches for $\mathbb{E}[u_t|\mathbf{b}_t]$. In practice, researchers may assume $\mathbb{E}[u_t|\mathbf{b}_t] = 0$ or, alternatively, employ flexible-parametric/nonparametric approaches to form the estimator $\widehat{\mathbb{E}}[u_t|\mathbf{b}_t]$. Since the latter approaches are setting-specific and technically well-understood, we chose to present our results using the generic condition (SA-1.v) rather than providing primitive conditions for a specific example of $\widehat{\mathbb{E}}[u_t|\mathbf{b}_t]$.

4.1.2. Multi-Equation With Weakly Dependent Data

The second example is the multi-equation setup introduced in Section 2, where we incorporate pre-intervention covariates in the construction of the SC weights and allow for stationary weakly dependent time series data. See Kilian and Lütkepohl (2017) and references therein for an introduction to time series analysis. We let $M = 2$ (two features) and $K = 0$ (no additional controls) for simplicity, which gives the working model

$$a_{t,1} = \sum_{j=1}^{J} b_{jt,1}w_{0,j} + u_{t,1},$$

$$a_{t,2} = \sum_{j=1}^{J} b_{jt,2} w_{0,j} + u_{t,2},$$

$t = 1, \dots, T_0$. The first equation could naturally correspond to pre-intervention outcomes as in the previous example, that is, $a_{t,1} := Y_{1t}(0)$ and $\mathbf{b}_{t,1} := (Y_{2t}(0), Y_{3t}(0), \dots, Y_{(N+1)t}(0))'$, while the second equation could correspond to some other covariate (such as population density in the Basque terrorism application) also used to construct $\widehat{\mathbf{w}}$ in Equation (4). Let $\mathbf{b}_{t,l} = (b_{1t,l}, \cdots, b_{Jt,l})'$, for $l = 1, 2$. To provide interpretable primitive conditions, we also assume $\mathbf{u}_t = (u_{t,1}, u_{t,2})'$ and $\mathbf{b}_t = (\mathbf{b}_{t,1}', \mathbf{b}_{t,2}')'$ follow independent first-order stationary autoregressive (AR) processes:

$$\mathbf{u}_t = \mathbf{H}_u \mathbf{u}_{t-1} + \boldsymbol{\zeta}_{t,u}, \qquad \mathbf{H}_u = \operatorname{diag}(\rho_{1,u}, \rho_{2,u}),$$
$$\mathbf{b}_t = \mathbf{H}_b \mathbf{b}_{t-1} + \boldsymbol{\zeta}_{t,b}, \qquad \mathbf{H}_b = \operatorname{diag}(\rho_{1,b}, \rho_{2,b}, \dots, \rho_{J,b}),$$

where $\boldsymbol{\zeta}_{t,u}$ and $\boldsymbol{\zeta}_{t,b}$ are iid over t, independent of each other, and $\operatorname{diag}(\cdot)$ denotes a diagonal matrix with the function arguments as the corresponding diagonal elements. Let $\mathbf{D} = T_0^{1/2}\mathbf{I}_d$, and note that $\mathbf{U} = (u_{1,1}, \dots, u_{T_0,1}, u_{1,2}, \dots, u_{T_0,2})'$ in this case. A natural, generic variance estimator is

$$\widehat{\Sigma} = \frac{1}{T_0}\mathbf{Z}'\widehat{\mathbb{V}}[\mathbf{U}|\mathscr{H}]\mathbf{Z},$$

where $\widehat{\mathbb{V}}[\mathbf{U}|\mathscr{H}]$ is an estimate of $\mathbb{V}[\mathbf{U}|\mathscr{H}]$. In this example, Σ corresponds to the (conditional) long-run variance, and naturally $\widehat{\Sigma}$ can be chosen to be any standard estimator thereof.

Theorem SA-2 in the supplemental appendix gives primitive conditions that verify the high-level conditions of Theorem A in the appendix, and the high-level conditions of Theorem 2 for implementation. Note that because of the time dependence in this example, the primitive conditions are for Theorem A instead of Theorem 1. In particular, we show that under standard conditions guaranteeing β-mixing and moment and rank conditions (similar to those imposed in the previous example), the conditions of Theorem A hold with $\pi_\gamma = \mathfrak{C}_\pi T_0^{-\mathfrak{c}_\pi}$ and $\epsilon_\gamma = \mathfrak{C}_\epsilon T_0^{-\mathfrak{c}_\epsilon}$ for nonnegative constants $\mathfrak{C}_\pi$ and $\mathfrak{C}_\epsilon$, and some positive constants $\mathfrak{c}_\pi$ and $\mathfrak{c}_\epsilon$, which are characterized precisely in the supplemental appendix. Theorem 2 is also verified using the primitive conditions imposed in Theorem SA-2 in the supplemental appendix, and the associated nonasymptotic constants are characterized in its proof.

As in the previous example, Theorem SA-2 in the supplemental appendix illustrates the kind of primitive conditions needed to quantify in-sample uncertainty using our proposed methods. In this example, we accommodate multiple covariates (equations) in the construction of the SC weights and also allow for AR(1) dependent (stationary) time series data. The only intentionally high-level condition imposed is (SA-2.v), $\mathbb{P}(\mathbb{P}(||\widehat{\Sigma} - \Sigma|| \leq \epsilon^\star_{\Sigma,1}|\mathscr{H}) \geq 1 - \epsilon^\star_{\Sigma,2}) \geq 1 - \pi^\star_\Sigma$ for nonnegative constants $\epsilon^\star_{\Sigma,1}$, $\epsilon^\star_{\Sigma,2}$ and $\pi^\star_\Sigma$, which requires a concentration probability bound for the long-run variance estimator $\widehat{\Sigma}$ used to approximate the quantiles of the conditional (on $\mathscr{H}$) distribution of the bounding random variables ς_{L} and ς_{U} via simulations (Theorem 2). This condition is not difficult to verify for specific examples.

4.1.3. Cointegration

Our third and final example illustrates how nonstationary data can also be handled within our framework. See Tanaka (2017) and references therein for an introduction to nonstationary time series analysis. Suppose that for each $1 \leq l \leq M$, $\{a_{t,l}\}_{t=1}^T$, $\{b_{1t,l}\}_{t=1}^T, \dots, \{b_{Jt,l}\}_{t=1}^T$ are $I(1)$ processes, and $\{c_{1t,l}\}_{t=1}^T, \dots, \{c_{Kt,l}\}_{t=1}^T$ and $\{u_{t,l}\}_{t=1}^T$ are $I(0)$ processes. Therefore, $\mathbf{A}$ and $\mathbf{B}$ form a cointegrated system. For simplicity, consider the following example: for each $l = 1, \dots, M$ and $j = 1, \dots, J$,

$$a_{t,l} = \sum_{j=1}^{J} b_{jt,l} w_{0,j} + \sum_{k=1}^{K} c_{kt,l} r_{0,k,l} + u_{t,l},$$
$$b_{jt,l} = b_{j(t-1),l} + v_{jt,l},$$

where $u_{t,l}$ and $v_{jt,l}$ are stationary unobserved disturbances. In this scenario, $(1, -\mathbf{w}_0')'$ plays the role of a cointegrating vector such that the linear combination of $\mathbf{A}$ and $\mathbf{B}$ is stationary. The normalizing matrix $\mathbf{D} = \operatorname{diag}\{T_0, \dots, T_0, \sqrt{T_0}, \dots, \sqrt{T_0}\}$, where the first J elements are T_0 and the remaining ones are $\sqrt{T_0}$. Let $\check{\mathbf{Z}}_t = (\check{\mathbf{z}}_{t,1}, \dots, \check{\mathbf{z}}_{t,M})$, where $\check{\mathbf{z}}_{t,l}$ is the $((l-1)T_0 + t)$th column of $\operatorname{diag}\{T_0^{-1/2}\mathbf{I}_J, \mathbf{I}_{KM}\}\mathbf{Z}'$, for $l = 1, \dots, M$. Recall that $\mathbf{u}_t = (u_{t,1}, \dots, u_{t,M})'$. Write $\mathbf{v}_{t,l} = (v_{1t,l}, \dots, v_{Jt,l})'$, $\mathbf{v}_t = (\mathbf{v}_{t,1}', \dots, \mathbf{v}_{t,M}')'$, and $\mathbf{c}_{t,l} = (c_{1t,l}, \dots, c_{kt,l})'$. We allow some elements in $\mathbf{v}_t$ to be used in $\{\mathbf{c}_{t,l}\}_{l=1}^M$. Let $\mathbf{q}_t$ collect all distinct variables in $\mathbf{u}_t, \mathbf{v}_t, \mathbf{c}_{t,1}, \cdots, \mathbf{c}_{t,M}$. As in the previous example, a generic variance estimator is

$$\widehat{\Sigma} = \frac{1}{T_0}\sum_{t=1}^{T_0} \check{\mathbf{Z}}_t \widehat{\mathbb{V}}[\mathbf{u}_t|\mathscr{H}]\check{\mathbf{Z}}_t',$$

where $\widehat{\mathbb{V}}[\mathbf{u}_t|\mathscr{H}]$ is an estimate of $\mathbb{V}[\mathbf{u}_t|\mathscr{H}]$.

Theorem SA-3 in the supplemental appendix gives more primitive conditions and verifies the high-level conditions of Theorems 1 and 2 in the cointegration scenario. More precisely, it provides conditions so that Theorem 1 holds with $\pi_\gamma = \mathfrak{C}_{\pi,1}T_0^{-\psi\nu} + \mathfrak{C}_{\pi,2}T_0^{-1} + \pi_{Q,1} + \pi_{Q,2}$ and $\epsilon_\gamma = \mathfrak{C}_\epsilon(\log T_0)^{\frac{3}{2}(1+\mathfrak{c}_Q)}T_0^{-1/2}$ for nonnegative constants $(\psi, \nu, \pi_{Q,1}, \pi_{Q,2}, \mathfrak{c}_Q)$ specified in the assumptions of the theorem and nonnegative constants $(\mathfrak{C}_{\pi,1}, \mathfrak{C}_{\pi,2}, \mathfrak{C}_\epsilon)$ characterized in the proof. Similarly, Theorem 2 is also verified under more primitive conditions, including a higher-level condition of the form $\mathbb{P}(\mathbb{P}(||\widehat{\Sigma} - \Sigma|| \leq \epsilon^\star_{\Sigma,1}|\mathscr{H}) \geq 1 - \epsilon^\star_{\Sigma,2}) \geq 1 - \pi^\star_\Sigma$ for nonnegative constants $\epsilon^\star_{\Sigma,1}$, $\epsilon^\star_{\Sigma,2}$ and $\pi^\star_\Sigma$, as in the previous examples.

When $\mathbf{C}$ is excluded, $\widehat{\mathbf{w}}$ is a least-square estimator of the cointegrating vector, which is typically biased due to the potential correlation between $\mathbf{v}_t$ and $\mathbf{u}_t$. In Theorem SA-3 in the supplemental appendix, we include $\mathbf{C}$ and allow it to include contemporary $\mathbf{v}_t$ to correct this bias. More generally, one may augment the regression with $\mathbf{v}_t$ and its leads and lags, which is termed dynamic OLS in the time series literature. The results for this general case may be established using a similar strategy.

5. Out-of-Sample Uncertainty

The unobserved random variable e_T in (7) is a single error term in period T, which we interpret as the error from out-of-sample

prediction, conditional on $\mathscr{H} = \sigma(\mathbf{B}, \mathbf{C}, \mathbf{x}_T, \mathbf{g}_T)$. Naturally, in order to set appropriate $M_{2,\mathrm{L}}$ and $M_{2,\mathrm{U}}$ in Lemma 1, it is necessary to determine certain features of the conditional distribution $\mathbb{P}[e_T \leq \cdot|\mathscr{H}]$. In turn, determining those features would require strong distributional assumptions between pretreatment and post-treatment periods, or perhaps across units. In this section we propose principled but agnostic approaches to quantify the uncertainty introduced by the post-treatment unobserved shock e_T. Since formalizing the validity of our methods requires strong assumptions, in this article we recommend a generic sensitivity analysis to incorporate out-of-sample uncertainty to the prediction intervals. In particular, we propose employing three distinct methods for quantifying the uncertainty introduced by e_T as a starting point, and then assessing more generally whether the additional uncertainty would render the prediction intervals large enough to eliminate any statistically significant treatment effect.

Our starting point is a nonasymptotic probability bound on e_T via concentration inequalities. Such textbook results can be found in, for example, Vershynin (2018) and Wainwright (2019). We rely on the following lemma, which provides the desired bounds for e_T under different moment-like conditions.

Lemma 3 (Non-Asymptotic Probability Concentration for e_T).

(G) If there exists some $\sigma_{\mathscr{H}} > 0$ such that $\mathbb{E}[\exp(\lambda(e_T - \mathbb{E}[e_T|\mathscr{H}]))|\mathscr{H}] \leq \exp(\sigma^2_{\mathscr{H}}\lambda^2/2)$ a.s. for all $\lambda \in \mathbb{R}$, then for any $\varepsilon > 0$, $\mathbb{P}(|e_T - \mathbb{E}[e_T|\mathscr{H}]| \geq \varepsilon|\mathscr{H}) \leq 2\exp(-\varepsilon^2/(2\sigma^2_{\mathscr{H}}))$.

(P) If $\mathbb{E}[|e_T|^m|\mathscr{H}] < \infty$ a.s. for some $m \geq 2$, then for any $\varepsilon > 0$, $\mathbb{P}(|e_T - \mathbb{E}[e_T|\mathscr{H}]| \geq \varepsilon|\mathscr{H}) \leq \varepsilon^{-m}\mathbb{E}[|e_T - \mathbb{E}[e_T|\mathscr{H}]|^m|\mathscr{H}]$.

This lemma gives (possibly crude) bounds on the necessary features of the conditional distribution of e_T given $\mathscr{H}$. Lemma 3(G) corresponds to a sub-Gaussian tail assumption, while Lemma 3(P) exploits only a polynomial bound on moments of $e_t|\mathscr{H}$. In both cases, the only unknowns are the "center" and "scale" of the distribution: $\mathbb{E}[e_T|\mathscr{H}]$ and $\sigma^2_{\mathscr{H}}$ (or higher-moments), respectively. These unknown features can be estimated or tabulated based on (i) model assumptions and (ii) observed pretreatment data, at least as an initial step toward a sensitivity analysis.

For practical purposes, we first outline three alternative strategies to assess the uncertainty coming from e_T, starting with Lemma 3 and progressively adding more restrictions. After introducing these approaches, we turn to discussing how they can be used as an initial step toward a principled sensitivity analysis for uncertainty quantification of the SC estimator. Section 6 illustrates this idea using simulated data and empirical applications.

- *Approach 1: Nonasymptotic Bounds*. In view of Lemma 3, we only need to extract some features of $e_T|\mathscr{H}$, namely some conditional moments of the form $\mathbb{E}[|e_T|^m|\mathscr{H}]$ (or $\mathbb{E}[e_T^m|\mathscr{H}]$) for appropriate choice(s) of $m \geq 1$. In practice, for example, pretreatment residuals $\{\widehat{u}_t\}_{t=1}^{T_0}$ could be used to estimate those quantities (e.g., under stationarity and other regularity conditions). Alternatively, the necessary conditional moments could be set using external information, or tabulated across different values to assess the sensitivity of the resulting prediction intervals. Importantly, once $\mathbb{E}[e_T|\mathscr{H}]$ and $\sigma^2_{\mathscr{H}}$ (or higher-moments) are set, then computing $M_{2,\mathrm{L}}$ and $M_{2,\mathrm{U}}$ in Lemma 1 is straightforward via Lemma 3.
- *Approach 2: Location-scale Model*. Suppose that $e_T = \mathbb{E}[e_T|\mathscr{H}] + (\mathbb{V}[e_T|\mathscr{H}])^{1/2}\varepsilon_T$ with ε_T statistically independent of $\mathscr{H}$. This setting imposes restrictions on the distribution of $e_T|\mathscr{H}$, but allows for a much simpler tabulation strategy. Specifically, the bounds in Lemma 1 can now be set as $M_{2,\mathrm{L}} = \mathbb{E}[e_T|\mathscr{H}] + (\mathbb{V}[e_T|\mathscr{H}])^{1/2}\mathfrak{c}_\varepsilon(\alpha_2/2)$ and $M_{2,\mathrm{U}} = \mathbb{E}[e_T|\mathscr{H}] + (\mathbb{V}[e_T|\mathscr{H}])^{1/2}\mathfrak{c}_\varepsilon(1-\alpha_2/2)$ where $\mathfrak{c}_\varepsilon(\alpha_2/2)$ and $\mathfrak{c}_\varepsilon(1-\alpha_2/2)$ are $\alpha_2/2$ and $(1-\alpha_2/2)$ quantiles of ε_t, respectively, and α_2 is the desired prespecified level. In practice, $\mathbb{E}[e_T|\mathscr{H}]$ and $\mathbb{V}[e_T|\mathscr{H}]$ can be parameterized and estimated using the pre-intervention residuals $\{\widehat{u}_t\}_{t=1}^{T_0}$, or perhaps tabulated using auxiliary information. Once such estimates are available, the appropriate quantiles can be easily obtained using the standardized (estimated) residuals. This approach is likely to deliver more precise prediction intervals when compared to Approach 1, but at the expense of potential misspecification due to the location-scale model used.
- *Approach 3: Quantile Regression*. In view of Lemma 1, we only need to determine the $\alpha_2/2$ and $(1-\alpha_2/2)$ conditional quantiles of $e_T|\mathscr{H}$. Consequently, another possibility is to employ quantile regression methods to estimate those quantities using pretreament data.

While the three approaches above are simple and intuitive, they are potentially unsatisfactory because their validity would require arguably strong assumptions on the underlying data generating process linking the pretreatment and post-treatment data. Such assumptions, however, are difficult to avoid because the ultimate goal is to learn about uncertainty introduced by an unobserved random variable after the treatment began (i.e., $e_T|\mathscr{H}$ for $T > T_0$). Without additional data availability or specific modeling assumptions allowing for transferring information from the pretreatment period into the post-treatment period, it is difficult to formally set $M_{2,\mathrm{L}}$ and $M_{2,\mathrm{U}}$ in Lemma 1.

Nevertheless, it is possible to approach out-of-sample uncertainty quantification as a principled sensitivity analysis, using the methods above as a starting point. Given the formal and detailed in-sample uncertainty quantification developed in the previous section, it is natural to progressively enlarge the final prediction intervals by adding additional out-of-sample uncertainty to then ask the question: how large does the additional out-of-sample uncertainty contribution coming from $e_T|\mathscr{H}$ need to be in order to render the treatment effect τ_t in Equation (1) statistically insignificant? Using the approaches above, or similar ones, it is possible to construct natural initial benchmarks. For instance, the variability displayed by the pretreatment outcomes or SC residuals can help guide the level of "reasonable" out-of-sample uncertainty. Alternatively, in specific applications, natural levels of uncertainty for the outcomes of interest could be available, and hence used to tabulate the additional out-of-sample uncertainty. In Section 6 we further discuss and illustrate this idea numerically.

5.1. Examples

We revisit the three examples considered in Section 4.1 and illustrate how the implementation of the three approaches outlined earlier may accommodate different assumptions on the data generating process. We discuss the outcomes-only case in more detail, which suffices to showcase our basic strategy of out-of-sample uncertainty quantification. For the other two examples, we briefly explain some important conceptual and implementational issues. As mentioned above, these methods rely on strong assumptions and should be viewed as a starting point of a general sensitivity analysis.

5.1.1. Outcomes-Only

Recall that the data is assumed to be iid over $1 \leq t \leq T$ in this case. The conditional distribution of e_T given $\mathscr{H}$ then reduces to that given the contemporary covariates $\mathbf{b}_T$ only. Also, we set $\mathbf{x}_T = \mathbf{b}_T$ and $\mathbf{g}_T = (1, \ldots, 1)'$. Then, the out-of-sample error e_T is equivalent to the pseudo-true residual u_T. By stationarity of the data, the information about the conditional distribution of u_T can be learned using the pre-intervention residuals. These substantial simplifications facilitate the implementation of the proposed methods for quantifying the out-of-sample uncertainty.

- *Approach 1: Nonasymptotic Bounds*. In general, we only need to estimate several conditional moments of u_T given $\mathbf{b}_T$. For example, assume that a (conditional) Gaussian bound holds for u_T. If $\mathbb{E}[u_T|\mathbf{b}_T] = 0$, that is, then the SC prediction correctly characterizes the conditional expectation of a_T given $\mathbf{b}_T$, then an estimate of the conditional variance of u_T suffices to construct a prediction interval for u_T. Otherwise, an estimate of $\mathbb{E}[u_T|\mathbf{b}_T]$ is also required to adjust the location of the prediction interval. These quantities can be estimated using the pretreatment data. Though the pseudo-true residuals $\{u_t\}_{t=1}^{T_0}$ are not observed, good proxies $\{\widehat{u}_t\}_{t=1}^{T_0}$ are available from the SC fitting. In practice, flexible parametric or nonparametric approaches can be used to estimate these conditional moments. For instance, we can implement a simple linear regression of $\widehat{u}_t$ on $\mathbf{b}_t$ to estimate $\mathbb{E}[u_t|\mathbf{b}_t]$. Denote the predicted values by $\widehat{\mathbb{E}}[u_t|\mathbf{b}_t]$. For the conditional variance, specify a model $\mathbb{V}[u_t|\mathbf{b}_t] = \exp(\mathbf{b}_t'\boldsymbol{\theta}_b + \theta_0)$ and implement a regression of $\log((\widehat{u}_t - \widehat{\mathbb{E}}[u_t|\mathbf{b}_t])^2)$ on $\mathbf{b}_t$. The predicted conditional variance is guaranteed to be positive. A prediction interval for the out-of-sample error can then be constructed based on Lemma 3(G).
- *Approach 2: Location-scale Model*. Similarly, to implement Approach 2, we only need estimates of the conditional mean and variance of u_T given $\mathbf{b}_T$, denoted by $\widehat{\mathbb{E}}[u_t|\mathbf{b}_t]$ and $\widehat{\mathbb{V}}[u_t|\mathbf{b}_t]$ respectively. They can be obtained using the methods outlined previously. Once they are available, set $M_{2,\mathrm{L}} = \widehat{\mathbb{E}}[u_T|\mathbf{b}_T] + (\widehat{\mathbb{V}}[u_T|\mathbf{b}_T])^{1/2}\widehat{\mathfrak{c}}_\varepsilon(\alpha_2/2)$ and $M_{2,\mathrm{U}} = \widehat{\mathbb{E}}[u_T|\mathbf{b}_T] + (\widehat{\mathbb{V}}[u_T|\mathbf{b}_T])^{1/2}\widehat{\mathfrak{c}}_\varepsilon(1-\alpha_2/2)$ where $\widehat{\mathfrak{c}}_\varepsilon(\alpha_2/2)$ and $\widehat{\mathfrak{c}}_\varepsilon(1-\alpha_2/2)$ are $\alpha_2/2$ and $(1-\alpha_2/2)$ quantiles of $\{\widehat{\varepsilon}_t\}_{t=1}^{T_0}$ where $\widehat{\varepsilon}_t = (\widehat{u}_t - \widehat{\mathbb{E}}[u_t|\mathbf{b}_t])/(\widehat{\mathbb{V}}[u_t|\mathbf{b}_t])^{1/2}$, respectively.
- *Approach 3: Quantile Regression*. We can estimate the $\alpha_2/2$ and $(1-\alpha_2/2)$ conditional quantiles of u_t given $\mathbf{b}_t$ parametrically or nonparametrically. For instance, assume the ℓth quantile of u_t admits a linear form: $Q(\ell|\mathbf{b}_t) = \mathbf{b}_t'\theta(\ell)$. Then, we can implement a quantile regression of the pretreatment residuals $\widehat{u}_t$ on $\mathbf{b}_t$ for $\ell = \alpha_2/2$ and $(1-\alpha_2/2)$, which suffices to construct a bound on e_T. See Koenker et al. (2017), and references therein, for a comprehensive discussion of quantile regression methods.

5.1.2. Multi-Equation with Weakly Dependent Data

When data is weakly dependent and multiple features are used in the construction of SC weights, the implementation of the three approaches is similar to that in the outcomes-only case, but two outstanding issues need to be addressed. First, as described in Section 4.1.2, the SC weights are obtained by matching on two pre-intervention features $\{a_{t,1}\}_{t=1}^{T_0}$ and $\{a_{t,2}\}_{t=1}^{T_0}$, while in most SC applications, the final counterfactual prediction is constructed by setting $\mathbf{x}_T = \mathbf{b}_T$ (and $\mathbf{g}_T = \mathbf{0}$ in this case). Conceptually, the out-of-sample error $e_T = Y_{1T}(0) - \mathbf{b}_T'\mathbf{w}_0$ may or may not correspond to the pseudo-true residual $\mathbf{u}_t$ prior to the treatment. For example, if the pretreatment outcomes are used in the first equation, then by construction, e_t in this scenario is equivalent to $u_{t,1}$ for $t = 1, \ldots, T$. In the pre-intervention period, the residuals $\{\widehat{u}_{t,1}\}_{t=1}^{T_0}$ from the SC fitting play the role of proxies for $\{u_{t,1}\}_{t=1}^{T_0}$. In contrast, if pretreatment outcomes are not used in any of the two equations, then e_t is generally not the same as $u_{t,1}$ or $u_{t,2}$. Nevertheless, we can manually construct $\widehat{e}_t = Y_{1t}(0) - \mathbf{b}_t'\widehat{\mathbf{w}}$ as a proxy for e_t in the pre-intervention period.

Second, the dependence of e_T on $\mathscr{H}$ should be appropriately characterized. Consider a simple scenario where pretreatment outcomes are used in the construction of SC weights so that $e_t = u_{t,1}$. By our assumptions on $\boldsymbol{\zeta}_{t,u}$ and $\boldsymbol{\zeta}_{t,b}$, $\{u_{t,1}\}_{t=1}^{T}$ is independent of $\{\mathbf{b}_t\}_{t=1}^{T}$. If we further assume the two components of $\boldsymbol{\zeta}_{t,u}$ are independent of each other, then the conditional distribution of $u_{T,1}$ given $\mathscr{H}$ reduces to its unconditional distribution. Therefore, to implement the three approaches, one only needs to estimate the unconditional mean, variance or quantiles using the residuals $\{\widehat{u}_{t,1}\}_{t=1}^{T_0}$. In practice, however, the independence between $u_{T,1}$ and $\mathscr{H}$ may be unrealistic. Assuming an appropriate weak dependence structure, we can still characterize or approximate the conditional mean, variance or quantiles of $u_{t,1}$ by functions of $\mathbf{b}_t$ and lags thereof, which could be estimated by parametric or nonparametric regressions using the pre-intervention data.

5.1.3. Cointegration

As in the second example, we first determine the pre-intervention analogue to the out-of-sample error e_T. For instance, we let $a_{t,1} = Y_{1t}(0)$ and $b_{jt,1} = Y_{(j+1)t}(0)$ for $j = 1, \ldots, N$. In practice, the final counterfactual prediction is often constructed by setting $\mathbf{x}_T = (Y_{2t}(0), \ldots, Y_{(N+1)t}(0))'$ and $\mathbf{g}_T = \mathbf{0}$, that is, no additional control variables are used in the out-of-sample prediction. Then, we have $e_t = u_{t,1} + \sum_{k=1}^{K} c_{kt,1} r_{0,k,1}$ for $t = 1, \ldots, T$. In view of the assumptions imposed in Theorem SA-3 in the supplemental appendix, the conditional distribution of e_t given $\mathscr{H}$ reduces to that given the contemporary variables $\{\mathbf{v}_t, \mathbf{c}_{t,1}, \ldots, \mathbf{c}_{t,M}\}$. As in the previous examples, we can estimate its conditional mean, variance or quantiles using various parametric or nonparametric methods.

The assumption that $\{\mathbf{q}_t\}_{t=1}^T$ is iid in Theorem SA-3 may be too strong. In practice, as mentioned previously, we may want to augment the regression of the residual $\widehat{e}_t$ by lags (and leads) of $\mathbf{v}_t$ and $\{\mathbf{c}_{t,l}\}_{l=1}^M$ and transformations thereof to take into account potential time series dependence.

6. Numerical Results

We illustrate the performance of the proposed prediction intervals with a Monte Carlo experiment and two empirical examples. To implement the methods described in Section 4 and 5, we take a simple "plug-in" estimator $\widehat{\Sigma}$ of the long-run variance Σ and employ parametric polynomial regressions to estimate the conditional mean, variance and quantiles of e_T given $\mathscr{H}$ whenever needed. In addition, the choice of the tuning parameter ϱ can be based on a bound implied by optimization. Specifically, since $\widehat{\boldsymbol{\delta}}$ must satisfy the basic inequality specified in the definition of $\mathcal{M}_\xi$ (see Lemma 2), we can construct a threshold ϱ for $\widehat{\boldsymbol{\beta}}$ based on some estimates of the variance of $\widehat{\boldsymbol{\gamma}}$ and the eigenvalues of $\widehat{\mathbf{Q}}$. More details are discussed below, and we also provide complete replication codes. Last but not least, in view of the small sample size in many SC applications, these practical choices play the role of a reasonable starting point for a principled sensitivity analysis, as illustrated in this section.

6.1. Simulations

We conduct a Monte Carlo investigation of the finite sample performance of our proposed methods. We consider the outcomes-only case where $M = 1$, $J = N$, $K = 0$, and only the outcome variable is used. Then, $\mathbf{A}_1 = (Y_{11}, Y_{12}, \ldots, Y_{1T_0})'$, $\mathbf{b}_{j,1} = (Y_{(j+1)1}, Y_{(j+1)2}, \ldots, Y_{(j+1)T_0})'$. We set $T_0 = 100$, $T_1 = 1$, and $N = 10$. We consider three data generating processes for $b_{jt} := b_{jt,1}$: $b_{jt} = \rho b_{j(t-1)} + v_{jt}$ with $\rho \in \{0, 0.5, 1\}$, and where $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})' \sim$ iid $\mathsf{N}(0, \mathbf{I}_N)$.

To examine the conditional coverage, we first generate a sample of $\{b_{jt} : 1 \le t \le T_0 + T_1, 1 \le j \le N\}$ using one of the three models. We set 5 evaluation points in the post-treatment period: $\tilde{b}_{1(T_0+1)} := b_{1(T_0+1)} + \mathsf{c} \cdot \mathsf{sd}(b_{1t})$, where $\mathsf{c} \in \{-1, -0.5, 0, 0.5, 1\}$ and $\mathsf{sd}(b_{1t})$ is the sample standard deviation of $\{b_{1t}\}_{t=1}^{T_0}$. In other words, we construct 5 designs by varying the value of the first conditioning variable in the last period. Taking each of them as given, we generate the treated unit $a_t := Y_{1t} = \mathbf{b}_t'\mathbf{w} + u_t$ by randomly drawing the error $u_t \sim$ iid $\mathsf{N}(0, 0.5)$ independent of $\{\mathbf{b}_t\}_{t=1}^T$. We set $\mathbf{w} = (0.3, 0.4, 0.3, 0, \cdots, 0)'$. By construction, $\mathbf{w}$ is exactly equivalent to the pseudo-true weight $\mathbf{w}_0$ defined in Equation (5) and satisfies the positivity and sum-to-one constraints in the standard SC. We consider 5000 simulated datasets. Note that the design $\{\mathbf{b}_t\}_{t=1}^{T_0}$ is fixed throughout the simulation study, and we only draw a new realization of the error term u_t at each repetition.

For comparison, we also investigate the unconditional coverage of the prediction intervals. The models for a_t and $\mathbf{b}_t$ are the same as described above, but in this case the design $\{\mathbf{b}_t\}_{t=1}^T$ is not fixed across the simulation study. Instead, at each repetition we randomly draw $\{(a_t, \mathbf{b}_t)\}_{t=1}^{T_0+1}$. Therefore, the coverage probability obtained in this alternative exercise is *unconditional*, that is, not conditional on a fixed design.

In the above construction, the error term u_t is independent of the design and has mean zero. To check the performance of the proposed method in models with misspecification error ($\mathbb{E}[u_t|\mathscr{H}] \neq 0$), we also consider several other data generating processes: for models with $\rho = 0$ and $\rho = 0.5$, we generate $u_t = 0.2b_{1t} + \zeta_t$, while for the model with $\rho = 1$, $u_t = 0.9(b_{1t} - b_{1(t-1)}) + \zeta_t$, where $\zeta_t \sim$ iid $\mathsf{N}(0, 0.5)$. Notably, in these misspecified models, the weight $\mathbf{w}$ defined previously is no longer equivalent to the pseudo-true weight $\mathbf{w}_0$.

We focus on several versions of our proposed prediction intervals for the counterfactual outcome $Y_{1T}(0)$ of the treated unit with 90% nominal (conditional) coverage probability: "M1" denotes the prediction interval based on assuming a Gaussian bound and applying the concentration inequality in Lemma 3(G) ("approach 1"); "M1-S" is the same as "M1" except that we increase the (estimated) conditional standard deviation $\widehat{\sigma}_{\mathscr{H}}$ by a factor of 2; "M2" denotes the prediction interval based on the location-scale model ("approach 2"); and "M3" denotes the prediction interval based on linear quantile regression ("approach 3"). For comparison, we also include the prediction interval "CONF," which is based on the conformal method developed in Chernozhukov, Wüthrich, and Zhu (2021b). In the supplemental appendix, we also report the performance of the prediction interval based on the cross-sectional permutation method proposed in Abadie, Diamond, and Hainmueller (2010).

As mentioned previously, we choose the tuning parameter ϱ based on the basic optimization inequality $\varrho = \widehat{\sigma}_u(\log T_0)^{\mathsf{c}}/(\min_{1\le j\le J}\widehat{\sigma}_{b_j}T_0^{1/2})$, where $\widehat{\sigma}_u$ is the estimated (unconditional) standard deviation of u_t, $\widehat{\sigma}_{b_j}$ is the estimated (unconditional) second moment of b_{jt}, and $\mathsf{c} = 1$ if $\rho = 1$ and $\mathsf{c} = 0.5$ if $\rho = 0$ or 0.5. This strategy accommodates the simple data generating processes considered in simulations, and it can be tailored to better suit the assumptions in more complex statistical models as well. In addition, we use polynomial regression to estimate various features of the conditional distribution of u_t given $\mathscr{H}$. To avoid overfitting, we only use a subset of the J control outcomes which have nonzero weights in $\widehat{\mathbf{w}}^\star$. Regarding the long-run variance Σ, we take a simple plug-in estimator $\widehat{\Sigma} = \mathbf{D}^{-1}\mathbf{Z}'\mathrm{diag}\{\tilde{u}_1^2, \ldots, \tilde{u}_{T_0}^2\}\mathbf{Z}\mathbf{D}^{-1}$, where $\tilde{u}_t = \widehat{u}_t - \widehat{\mathbb{E}}[u_t|\mathscr{H}]$ and $\widehat{\mathbb{E}}[u_t|\mathscr{H}]$ is the estimate of the conditional mean of u_t given $\mathscr{H}$.

Panels A and B of Table 1 summarize the results for models with and without misspecification error, respectively, where we use linear regression methods to estimate the conditional mean, variance or quantiles whenever needed. The proposed prediction intervals exhibit good coverage properties throughout different data generating processes and evaluation points, though they are conservative in several cases. In contrast, the actual coverage probability of conformal prediction intervals developed in Chernozhukov, Wüthrich, and Zhu (2021b) is lower than the target nominal level in general. See Section SA-3 of the supplement for additional simulation evidence.

6.2. Empirical Illustration

We showcase our methods by reanalyzing two empirical examples from the SC literature. The first example concerns the effect

Table 1. Simulation evidence, linear regression methods.

	Panel A: Models with misspecification error									
	M1		M1-S		M2		M3		CONF	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
$\rho = 0$										
Cond. 1	0.949	2.168	0.987	2.891	0.979	2.625	0.978	2.682	0.864	1.646
2	0.931	2.118	0.982	2.844	0.970	2.577	0.967	2.624	0.854	1.642
3	0.922	2.136	0.977	2.867	0.966	2.599	0.957	2.630	0.842	1.643
4	0.928	2.242	0.979	2.980	0.966	2.710	0.956	2.719	0.830	1.651
5	0.936	2.406	0.981	3.154	0.971	2.881	0.960	2.862	0.819	1.665
Uncond.	0.961	2.373	0.991	3.094	0.982	2.833	0.979	2.892	0.886	1.687
$\rho = 0.5$										
Cond. 1	0.954	2.423	0.986	3.140	0.980	2.876	0.976	2.928	0.854	1.680
2	0.965	2.488	0.990	3.203	0.984	2.939	0.982	2.992	0.865	1.691
3	0.973	2.579	0.992	3.296	0.988	3.031	0.986	3.078	0.875	1.706
4	0.980	2.694	0.993	3.414	0.990	3.149	0.989	3.183	0.887	1.729
5	0.983	2.830	0.995	3.556	0.992	3.289	0.988	3.305	0.896	1.756
Uncond.	0.962	2.387	0.990	3.106	0.981	2.846	0.983	2.921	0.882	1.695
$\rho = 1$										
Cond. 1	0.979	2.798	0.995	3.602	0.991	3.308	0.979	3.278	0.896	3.403
2	0.978	2.730	0.995	3.554	0.990	3.252	0.987	3.270	0.985	3.369
3	0.970	2.756	0.992	3.637	0.987	3.317	0.982	3.287	0.170	3.345
4	0.956	2.896	0.982	3.876	0.975	3.525	0.951	3.331	0.000	3.333
5	0.935	3.188	0.969	4.323	0.961	3.926	0.901	3.413	0.000	3.336
Uncond.	0.974	2.970	0.994	3.733	0.990	3.458	0.989	3.530	0.895	3.443
	Panel B: Models without misspecification error									
	M1		M1-S		M2		M3		CONF	
	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
$\rho = 0$										
Cond. 1	0.943	2.154	0.985	2.871	0.976	2.609	0.973	2.661	0.879	1.619
2	0.927	2.105	0.980	2.826	0.966	2.563	0.962	2.604	0.880	1.617
3	0.918	2.125	0.974	2.852	0.960	2.587	0.953	2.612	0.880	1.621
4	0.922	2.231	0.975	2.966	0.961	2.699	0.953	2.701	0.881	1.632
5	0.937	2.394	0.978	3.139	0.966	2.868	0.955	2.841	0.881	1.649
Uncond.	0.960	2.358	0.990	3.073	0.981	2.810	0.981	2.878	0.888	1.656
$\rho = 0.5$										
Cond. 1	0.959	2.416	0.988	3.127	0.981	2.866	0.978	2.918	0.875	1.666
2	0.970	2.480	0.990	3.191	0.985	2.929	0.984	2.980	0.876	1.671
3	0.975	2.571	0.993	3.283	0.989	3.022	0.987	3.064	0.879	1.682
4	0.981	2.685	0.995	3.401	0.990	3.139	0.987	3.168	0.885	1.699
5	0.985	2.820	0.996	3.542	0.992	3.278	0.989	3.287	0.888	1.721
Uncond.	0.961	2.370	0.991	3.083	0.981	2.825	0.983	2.894	0.884	1.656
$\rho = 1$										
Cond. 1	0.964	2.533	0.988	3.278	0.983	3.006	0.972	2.972	0.860	1.561
2	0.963	2.406	0.990	3.134	0.983	2.867	0.979	2.892	0.851	1.548
3	0.947	2.364	0.983	3.106	0.972	2.834	0.961	2.835	0.854	1.542
4	0.923	2.411	0.965	3.200	0.955	2.915	0.918	2.801	0.859	1.540
5	0.887	2.571	0.940	3.447	0.927	3.136	0.848	2.800	0.847	1.547
Uncond.	0.982	2.773	0.997	3.485	0.993	3.227	0.991	3.287	0.868	1.642

Notes: Conditional mean, variance, and quantiles of u_t are estimated based on linear regression methods. CP = coverage probability, AL = average length. "M1": prediction interval for $Y_{1T}(0)$ based on the Gaussian concentration inequality with 90% nominal coverage probability; "M1-S": the same as "M1", but the estimated standard deviation is doubled in the construction; "M2": prediction interval for $Y_{1T}(0)$ based on the location-scale model with 90% nominal coverage probability; "M3": prediction interval for $Y_{1T}(0)$ based on quantile regression with 90% nominal coverage probability; "CONF" prediction interval for $Y_{1T}(0)$ based on the conformal method developed in Chernozhukov, Wüthrich, and Zhu (2021b) with 90% nominal coverage probability.

of California's tobacco control program, known as Proposition 99, on per capita cigarette sales (see Abadie, Diamond, and Hainmueller 2010, for more details). The second example corresponds to the economic impact of 1990 German reunification on West Germany (see Abadie 2021, for more details). To conserve space, the results for the second example are reported in Section SA-4 of the supplement.

The outcome variable of interest is per capita cigarette sales in California, which is arguably nonstationary. We consider both the raw data and the first-differenced data, corresponding to the analysis of levels and growth rates of per capita sales, respectively. In each scenario, we construct (i) the synthetic control prediction $\mathbf{x}_T'\widehat{\mathbf{w}}$ as commonly done in the literature; (ii) the prediction interval for the (conditionally non-random) "SC component" $\mathbf{x}_T'\mathbf{w}_0$ only; and (iii) three distinct prediction intervals for the counterfactual $Y_{1T}(0)$. More specifically, the prediction intervals for $Y_{1T}(0)$ are implemented using the three methods outlined in Section 5: (i) conditional subgaussian bound using Lemma 3(G), labeled as *approach 1*; (ii) conditional bound based on location-scale model, labeled as *approach 2*; (iii) conditional bound using conditional quantile regression of residuals, labeled as *approach 3*.

The three methods used to quantify the out-of-sample uncertainty can be viewed as particular instances of a more general sensitivity analysis. In other words, varying the additional uncertainty contribution of e_T in a principled way, researchers

can better understand its impact on the construction of the prediction intervals. We also illustrate this approach ("sensitivity analysis"): focusing on approach 1 for concreteness, we rely on Gaussian bounds in Lemma 3(G) to assess how the prediction intervals behave as the variance of e_T varies.

Accordingly, we present six plots: the SC prediction $\mathbf{x}_T'\widehat{\mathbf{w}}$, the prediction interval (PI) only for the synthetic unit $\mathbf{x}_T'\mathbf{w}_0$ with at least 95% nominal coverage probability, the three different constructions of PIs for the counterfactual $Y_{1T}(0)$ with at least 90% nominal coverage probability, and a sensitivity analysis for one chosen post-treatment period. Because the size of the donor pool is larger than the number of available pretreatment periods, our procedure may rely on loose bounds for Gaussian approximation errors in high-dimensional settings.

We first consider the raw data of per capita cigarette sales. Figure 1(a) shows the trajectory of per capita sales of the synthetic California (dashed blue) and the actual California (solid black). After 1988, the synthetic California series is above the observed one, suggesting a negative shock of Proposition 99 on cigarette sales in California. Figure 1(b) adds a 95% conservative prediction interval for the SC component of California that takes into account the in-sample uncertainty due to the estimated SC weights. We add the uncertainty associated with e_T in Figures 1(c)–(e). The observed sequence is generally below the prediction intervals for the counterfactual outcome of California, suggesting statistically significant effects of Proposition 99. Figure 1(f) shows the sensitivity analysis of the effect in 1989. The result is robust: the corresponding PIs are well separated from the observed outcome of California if we vary the estimated (conditional) standard deviation of e_T in a relatively wide range.

We also analyze the (log) growth rate of per capita cigarette sales. The result is reported in Figure 2. We can see that the observed growth rate during the post-treatment period is generally lower than the SC prediction, but throughout the three constructions of PIs for the counterfactual outcome, the observed series is within the PIs for most post-treatment periods except in 1989. These empirical findings suggest a statistical significant effect of the tobacco control program on the growth rate of per capita cigarette sales only in year 1989. The sensitivity analysis in Figure 2(f) shows that the significance of the effect in 1989 is robust.

7. Conclusion

The SC method is part of the standard program evaluation toolkit. Despite its popularity, many important methodological and theoretical developments remain outstanding. We focus on quantifying the uncertainty of the SC method in predicting the main quantity of interest, $\tau_T = Y_{1T}(1) - Y_{1T}(0)$, in the standard SC framework. This quantity is the difference between the observed outcome of the treated unit in a post-treatment period T, and the outcome that the treated unit would have had in the same period in the absence of treatment. Because we view τ_T as a random variable and there is a single treated unit, we propose conditional prediction intervals that offer finite-sample probability guarantees regarding the realization of the counterfactual treated outcome. Our approach takes the SC constrained least-square optimization approach as the starting point. We model the counterfactual of the treated unit in period T as the weighted sum of the untreated units' features at T (with weights estimated with pretreatment data), and an error term. This decomposition highlights two sources of uncertainty, one from the in-sample estimation of the SC weights in the pretreatment period, and the other from the post-treatment error that arises due to the unavoidable out-of-sample prediction involved in the SC method, which may include potential misspecification errors from the SC weights. Using finite-sample concentration bounds, we derive prediction intervals that incorporate both sources of uncertainty. Because the uncertainty stemming from the out-of-sample post-treatment error term is hard to handle (especially under general misspecification), we recommend combining the prediction interval for the SC outcome with a principled sensitivity analysis for the post-treatment error. Our empirical illustrations show that our methods perform well using both simulated and real data. A general-purpose software package is underway (Cattaneo et al. 2021).

Appendix A: Extension to Weakly Dependent Data

We generalize Theorem 1 to allow for β-mixing data. For $\mathbf{u}_t = (u_{t,1},\dots,u_{t,M})'$, define the (conditional on $\mathscr{H}$) mixing coefficient $\mathfrak{b}(\cdot;\mathscr{H})$ by

$$\mathfrak{b}(k;\mathscr{H}) = \max_{1\le l\le n-k} \tfrac{1}{2} \sup \Big\{ \sum_i \sum_j \Big|\mathbb{P}(\mathcal{E}_i\cap\mathcal{E}_j'|\mathscr{H}) - \mathbb{P}(\mathcal{E}_i|\mathscr{H})\mathbb{P}(\mathcal{E}_j'|\mathscr{H})\Big| : \{\mathcal{E}_i\} \text{ is a finite partition of } \sigma(\mathbf{u}_1,\dots,\mathbf{u}_l), \{\mathcal{E}_j'\} \text{ is a finite partition of } \sigma(\mathbf{u}_{l+k},\dots,\mathbf{u}_{T_0})\Big\}.$$

See Pham and Tran (1985) and Doukhan (2012) for properties and examples of mixing conditions.

Theorem A below combines a coupling result for dependent data (Berbee 1987), a Berry-Esseen bound for convex sets (Raič 2019), and results on anti-concentration of the Gaussian measure for convex sets (Chernozhukov, Chetverikov, and Kato 2015; Chernozhukov et al. 2017), together with the standard "small-block and large-block" technique. Decompose the sequence $\{1,\dots,T_0\}$ into "large" and "small" blocks: $\mathcal{J}_1 = \{1,\dots,q\}$, $\mathcal{J}_1' = \{q+1,\dots,q+v\}$, ..., $\mathcal{J}_m = \{(q+v)(m-1)+1,\dots,(q+v)(m-1)+q\}$, $\mathcal{J}_m' = \{(q+v)(m-1)+q+1,\dots,(q+v)m\}$, $\mathcal{J}_{m+1}' = \{(q+v)m+1,\dots,T_0\}$ where $m = \lfloor T_0/(q+v)\rfloor$, $q > v$ and $q+v \le T_0/2$, and where $\lfloor\cdot\rfloor$ denotes the floor operator. The parameters q and v, which depend on T_0, control the sizes of the large and small blocks, respectively, and will satisfy certain conditions in the theorem below. To simplify notation, we let $\mathbf{s}_t = (s_{1t},\dots,s_{dt})'$ be the summand in $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$ corresponding to time t, and define $\mathbf{S}_{k,\square} = \sum_{t\in\mathcal{J}_k}\mathbf{s}_t$ and $\mathbf{S}_{k,\diamond} = \sum_{t\in\mathcal{J}_k'}\mathbf{s}_t$. Accordingly, let $S_{jk,\square}$ and $S_{jk,\diamond}$ be the jth elements of $\mathbf{S}_{k,\square}$ and $\mathbf{S}_{k,\diamond}$, respectively. Let $\Sigma_\square = \sum_{k=1}^m \mathbb{V}[\mathbf{S}_{k,\square}|\mathscr{H}]$ and introduce

$$\bar\sigma^2(q) := \max_{1\le j\le d}\frac{1}{m}\sum_{k=1}^m \mathbb{V}\Big[q^{-1/2}\sum_{t\in\mathcal{J}_k} s_{jt}\Big|\mathscr{H}\Big],$$

$$\bar\sigma^2(v) := \max_{1\le j\le d}\frac{1}{m}\sum_{k=1}^m \mathbb{V}\Big[v^{-1/2}\sum_{t\in\mathcal{J}_k'} s_{jt}\Big|\mathscr{H}\Big].$$

Theorem A (Distributional Approximation, Dependent Case). Assume $\mathcal{W}$ and $\mathcal{R}$ are convex, and $\widehat{\boldsymbol{\beta}}$ in Equation (4) and $\boldsymbol{\beta}_0$ in Equation (5) exist. Let $\psi \ge 3$. In addition, for nonnegative finite constants η_1, $\bar\sigma$, $\pi_{\gamma,1}$, η_2, $\pi_{\gamma,2}$, η_3, $\pi_{\gamma,3}$, η_4, $\pi_{\gamma,4}$, η_5, $\pi_{\gamma,5}$, and η_6, the following conditions hold:

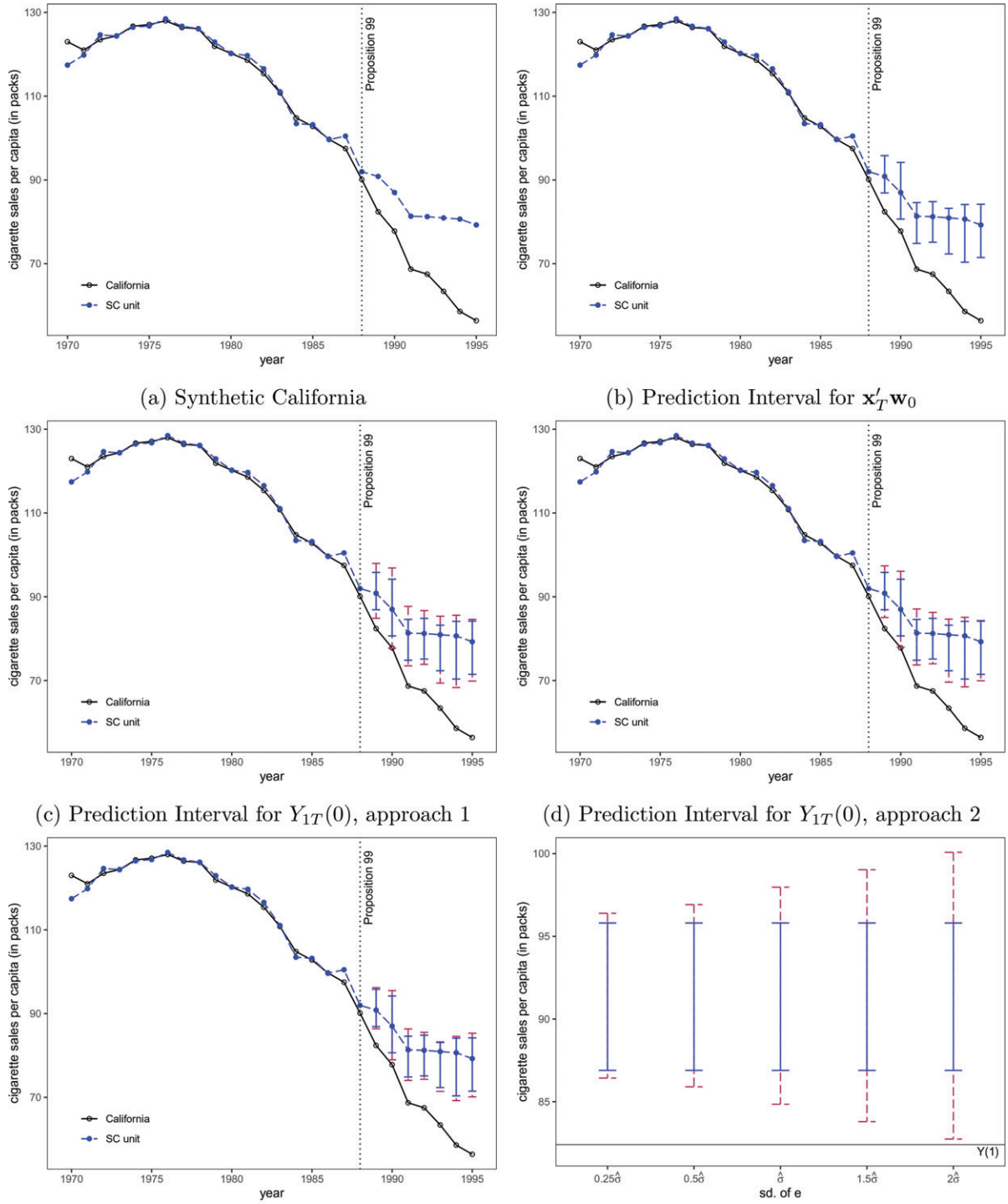

Figure 1. California tobacco control: cigarette sales per capita.
Notes. Panel (a): Per capita cigarette sales in California and synthetic California. Panel (b): Prediction interval for synthetic California with at least 95% coverage probability. Panels (c)–(e): Prediction interval for the counterfactual of California for with at least 90% coverage probability based on three methods described in Section 5, respectively. Panel (f): Prediction intervals for the counterfactual California based on approach 1, corresponding to $c \times \sigma_{\mathscr{H}}$, where $c = 0.25, 0.5, 1, 1.5, 2$. The horizontal solid line represents the observed outcome for the treated.

(TA.i) $\mathbf{u}_t$ is β-mixing conditional on $\mathscr{H}$ with mixing coefficient $\mathfrak{b}(\cdot;\mathscr{H})$;

(TA.ii) $\mathbb{P}\big(\mathbb{E}[\sum_{j=1}^{d}\sum_{k=1}^{m}|S_{jk,\diamond}|^{\psi}|\mathscr{H}] \le \eta_1,\ \bar{\sigma}^2(v) \le \bar{\sigma}^2\big) \ge 1 - \pi_{\gamma,1}$;

(TA.iii) $\mathbb{P}\big(\max_{1\le j\le d}\mathbb{E}[|S_{j(m+1),\diamond}|^{\psi}|\mathscr{H}] \le \eta_2\big) \ge 1 - \pi_{\gamma,2}$;

(TA.iv) $\mathbb{P}\big(\sum_{k=1}^{m}\mathbb{E}[||\Sigma_{\square}^{-1/2}\mathbf{S}_{k,\square}||^3|\mathscr{H}] \le \eta_3(42(d^{1/4}+16))^{-1}\big) \ge 1 - \pi_{\gamma,3}$;

(TA.v) $\mathbb{P}\big(||\Sigma_{\square}^{-1}||_{\mathrm{F}} \le d\eta_4\big) \ge 1 - \pi_{\gamma,4}$;

(TA.vi) $\mathbb{P}\big(||\Sigma_{\square}^{-1/2}\Sigma\Sigma_{\square}^{-1/2} - \mathbf{I}_d||_{\mathrm{F}} \le 2\eta_5\big) \ge 1 - \pi_{\gamma,5}$;

(TA.vii) $\max\big\{\eta_3, \eta_5, d\eta_4[\eta_6^{-1}(\sqrt{mv\bar{\sigma}^2\log d}+\eta_1^{1/\psi}\log d)+(d\eta_2)^{1/\psi}\eta_6^{-1/\psi}]\big\} \le \eta_6$ and $m\mathfrak{b}(v;\mathscr{H}) \le \eta_6$ a.s. on $\mathscr{H}$.

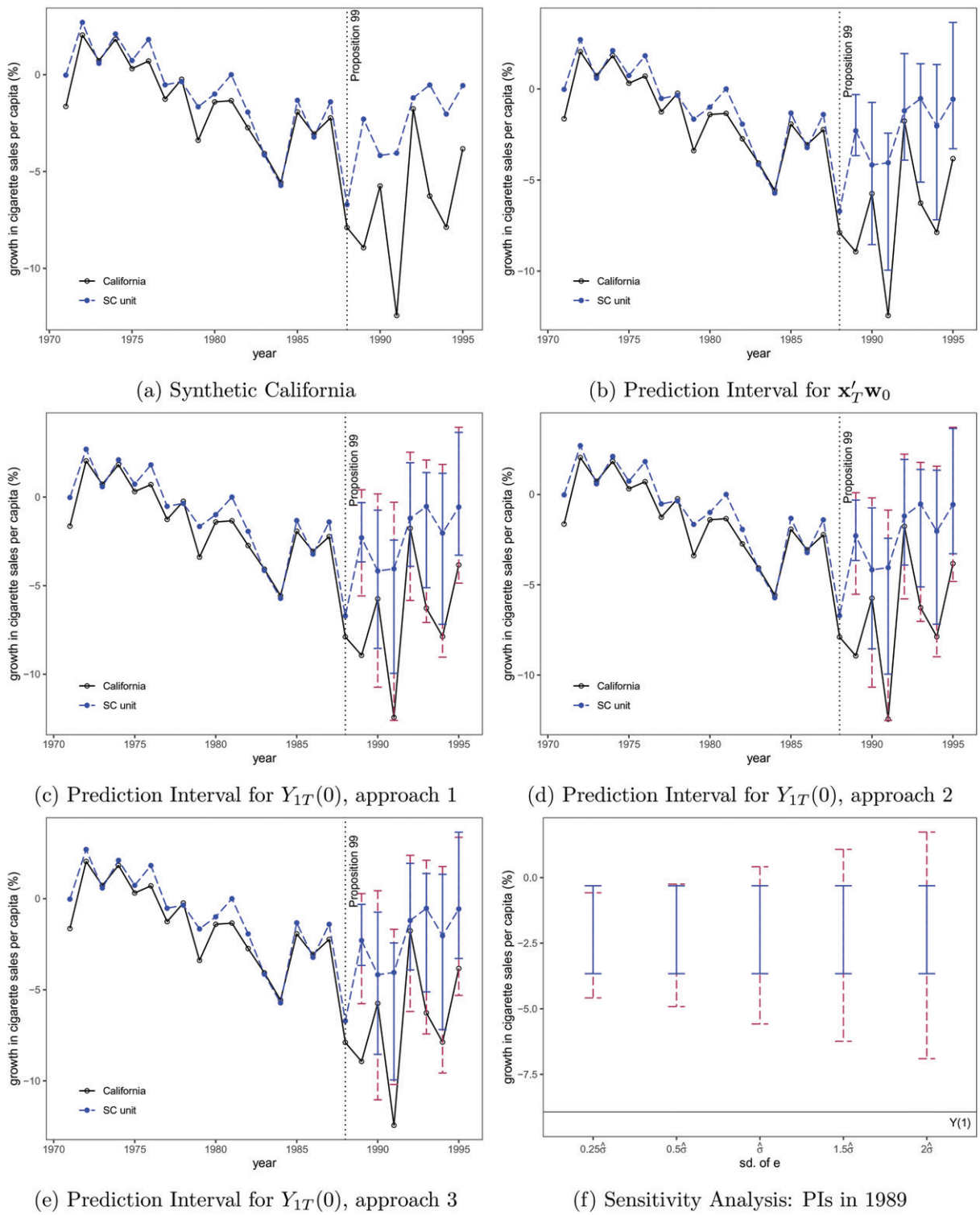

Figure 2. California tobacco control: growth rate of cigarette sales per capita.
Notes. Panel (a): Growth rate of per capita cigarette sales in California and synthetic California. Panel (b): Prediction interval for synthetic California with at least 95% coverage probability. Panels (c)-(e): Prediction interval for the counterfactual of California for with at least 90% coverage probability based on three methods described in Section 5, respectively. Panel (f): Prediction intervals for the counterfactual California based on approach 1, corresponding to $c \times \sigma_{\mathscr{H}}$, where $c = 0.25, 0.5, 1, 1.5, 2$. The horizontal solid line represents the observed outcome for the treated.

Then, for $\eta_5 \in [0, 1/4]$,

$$\mathbb{P}\Big[\mathbb{P}\big(\mathbf{p}_T'\mathbf{D}^{-1}\widehat{\boldsymbol{\delta}} \le \mathfrak{c}^{\dagger}(1-\alpha)\big|\mathscr{H}\big) \ge 1-\alpha-\epsilon_{\gamma}\Big] \ge 1-\pi_{\gamma},$$

where $\epsilon_{\gamma} = \mathfrak{C}\eta_6$ for finite positive constant $\mathfrak{C}$, which is characterized in the proof, $\pi_{\gamma} = \sum_{l=1}^{5}\pi_{\gamma,l}$, and $\mathfrak{c}^{\dagger}(1-\alpha)$ is the $(1-\alpha)$-quantile of $\varsigma_{\mathrm{U}}^{\dagger} = \sup\{\mathbf{p}_T'\mathbf{D}^{-1}\boldsymbol{\delta} : \boldsymbol{\delta} \in \mathcal{M}_{\mathbf{G}}\}$ conditional on $\mathscr{H}$.

Conditions (TA.i) and (TA.iv) are comparable to (T1.i) and (T2.i) in Theorem 1, respectively. The (conditional) independence assumption is relaxed to (conditional) β-mixing, and a bound on the conditional third moment of large blocks is imposed. The other conditions in Theorem A are new, and they ensure the small blocks and the last block can be neglected in a proper probability concentration sense.

Supplementary Materials

The online supplemental appendix contains theoretical proofs of the theorems, further details on the three examples presented, additional numerical evidence, and more discussion on the interpretation of the pseudo-true values underlying our procedure in specific settings.

Acknowledgments

We thank to Alberto Abadie, Amir Ali Ahmadi, Michael Jansson, Filippo Palomba and Mykhaylo Shkolnikov for many insightful discussions, and Kaspar Wüthrich for sharing replication codes used in the simulations.

Funding

We also thank to three reviewers for their critical comments and suggestions that helped improve our manuscript. Cattaneo and Titiunik gratefully acknowledges financial support from the National Science Foundation (SES-2019432), and Cattaneo gratefully acknowledges financial support from the National Institutes of Health (R01 GM072611-16).

References

Abadie, A. (2021), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59, 391–425. [1865,1876]

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503. [1865]

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105, 493–505. [1865,1868,1875,1876]

Abadie, A., and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. [1865,1867]

Abadie, A., and L'Hour, J. (2021), "A Penalized Synthetic Control Estimator for Disaggregated Data," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1971535. [1865]

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021), "On Robustness of Principal Component Regression," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1928513. [1865]

Amjad, M., Shah, D., and Shen, D. (2018), "Robust Synthetic Control," *The Journal of Machine Learning Research*, 19, 802–852. [1868]

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021), "Synthetic Difference in Differences," *American Economic Review*. [1868]

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021), "Matrix Completion Methods for Causal Panel Data Models," *Journal of the American Statistical Association*. doi: 10.3386/w25132. [1865]

Bai, J., and Ng, S. (2021), "Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1967163. [1865]

Ben-Michael, E., Feller, A., and Rothstein, J. (2021), "The Augmented Synthetic Control Method," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1929245. [1865]

Berbee, H. (1987), "Convergence Rates in the Strong Law for Bounded Mixing Sequences," *Probability Theory and Related Fields*, 74, 255–270. [1877]

Cattaneo, M., Feng, Y., Palomba, F., and Titiunik, R. (2021), "`scpi`: Uncertainty Quantification for Synthetic Control Estimators," Working Paper, Princeton University. [1866,1877]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), "Comparison and Anti-Concentration Bounds for Maxima of Gaussian Random Vectors," *Probability Theory and Related Fields*, 162, 47–70. [1877]

Chernozhukov, V., Chetverikov, D., Kato, K., et al. (2017), "Central Limit Theorems and Bootstrap in High Dimensions," *Annals of Probability*, 45, 2309–2352. [1877]

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021a), "Distributional Conformal Prediction," arXiv:1909.07889. [1866,1869]

——— (2021b), "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1920957. [1865,1868,1869,1875,1876]

——— (2021c), "A t-test for Synthetic Controls," arXiv:1812.10820. [1865]

Doudchenko, N., and Imbens, G. W. (2016), "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis," NBER Working Paper No. w22791. [1868]

Doukhan, P. (2012), *Mixing: Properties and Examples*, New York: Springer. [1870,1877]

Feng, Y. (2021), "Causal Inference in Possibly Nonlinear Factor Models," arXiv:2008.13651. [1865]

Ferman, B. (2021), "On the Properties of the Synthetic Control Estimator With Many Periods and Many Controls," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1965613. [1865]

Ferman, B., and Pinto, C. (2021), "Synthetic Controls with Imperfect Pre-Treatment Fit," *Quantitative Economics*. [1868]

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012), "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong With Mainland China," *Journal of Applied Econometrics*, 27, 705–740. [1868]

Kellogg, M., Mogstad, M., Pouliot, G., and Torgovitsky, A. (2021), "Combining Matching and Synthetic Controls to Trade off Biases from Extrapolation and Interpolation," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1979562. [1865]

Kilian, L., and Lütkepohl, H. (2017), *Structural Vector Autoregressive Analysis*, Cambridge: Cambridge University Press. [1871]

Koenker, R., Chernozhukov, V., He, X., and Peng, L. (2017), *Handbook of Quantile Regression*, Boca Raton, FL: CRC Press. [1874]

Li, K. T. (2020), "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*, 115, 2068–2083. [1865]

Masini, R., and Medeiros, M. C. (2021), "Counterfactual Analysis with Artificial Controls: Inference, High Dimensions and Nonstationarity," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1964978. [1865]

Pham, T. D., and Tran, L. T. (1985), "Some Mixing Properties of Time Series Models," *Stochastic Processes and Their Applications*, 19, 297–303. [1877]

Raič, M. (2019), "A Multivariate Berry–Esseen Theorem With Explicit Constants," *Bernoulli*, 25, 2824–2853. [1870,1877]

Shaikh, A. M., and Toulis, P. (2021), "Randomization Tests in Observational Studies With Staggered Adoption of Treatment," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2021.1974458. [1865]

Tanaka, K. (2017), *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory*, Hoboken, NJ: John Wiley & Sons. [1872]

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge: Cambridge University Press. [1866,1873]

Vovk, V. (2012), "Conditional Validity of Inductive Conformal Predictors," JMLR: Workshop and Conference Proceedings 25, Asian Conference on Machine Learning, pp. 475–490. [1866,1869]

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge: Cambridge University Press. [1866,1873]