

Distributionally Robust Synthetic Control: Ensuring Robustness Against Highly Correlated Controls and Weight Shifts

Taehyeon Koo¹ and Zijian Guo^{*2}

¹Columbia University Mailman School of Public Health

²Center for Data Science, Zhejiang University

November 5, 2025

Abstract

The synthetic control method estimates the causal effect by comparing the outcomes of a treated unit to a weighted average of control units that closely match the pre-treatment outcomes of the treated unit. This method presumes that the relationship between the potential outcomes of the treated and control units remains consistent before and after treatment. However, the estimator may become unreliable when these relationships shift or when control units are highly correlated. To address these challenges, we introduce the **D**istributionally **R**obust **S**ynthetic **C**ontrol (DRoSC) method by accommodating potential shifts in relationships and addressing high correlations among control units. The DRoSC method targets a new causal estimand defined as the optimizer of a worst-case optimization problem that checks through all possible synthetic weights that comply with the pre-treatment period. When the identification conditions for the classical synthetic control method hold, the DRoSC method targets the same causal effect as the synthetic control. When these conditions are violated, we show that this new causal estimand is a conservative proxy of the non-identifiable causal effect. We further show that the limiting distribution of the DRoSC estimator is non-normal and propose a novel inferential approach to characterize this non-normal limiting distribution. We demonstrate its finite-sample performance through numerical studies and an analysis of the economic impact of terrorism in the Basque Country.

1 Introduction

The synthetic control (SC) method (Abadie and Gardeazabal, 2003; Abadie et al., 2010, 2015) plays an increasingly important role in empirical research in economics and the social sciences, primarily due to its transparent construction and interpretability. The method estimates the counterfactual outcome for a treated unit by constructing a weighted average of control units, with weights chosen to closely match the pre-treatment trajectory of the treated unit. This synthetic control serves as an approximation of the potential outcome in the absence of treatment, enabling causal effect estimation through comparison with the observed post-treatment outcome. The SC framework has inspired a wide range of methodological developments and is recognized as a key contribution to the policy evaluation literature (Athey and Imbens, 2017).

^{*}Correspondence to Zijian Guo (zijguo@zju.edu.cn).

Despite its widespread adoption, the SC method faces limitations that may compromise its empirical reliability. One key challenge is the instability of learning the synthetic weights: in the presence of strong correlations among control units, multiple weight configurations may yield comparable pre-treatment fits, leading to instability in counterfactual predictions and, consequently, treatment effect estimates. A second issue is weight shift, that is, changes in the treated-control relationship between the pre- and post-treatment periods. Such weight shifts can invalidate their post-treatment application and induce bias in causal estimation.

When either challenge arises, the treatment effect is no longer point-identifiable. To address this, we define a new causal estimand through the lens of distributionally robust optimization (DRO). When neither of the aforementioned issues, such as high correlations among control units or weight shifts, occurs, we show that this estimand coincides with the average treatment effect for the treated unit. In contrast, when either challenge is present, the proposed estimand provides a conservative lower bound while preserving the sign of the treatment effect.

1.1 Our Results and Contributions

In this paper, we introduce a novel method—the Distributionally Robust Synthetic Control (DRoSC) estimator—which targets a new causal estimand defined as the solution to a distributionally robust optimization (DRO) problem within the SC framework. We refer to this new causal estimand as the *weight-robust treatment effect*. Crucially, this estimand remains identifiable even when the true treatment effect is not. Rather than assuming a uniquely identifiable post-treatment weight scheme, as in the standard SC framework, we consider a class of plausible weight configurations that may arise due to weight shifts or high correlations among control units. We then define a worst-case risk over this class and define the robust causal estimand as the treatment effect minimizing this risk. This formulation yields an estimand that remains meaningful even when standard SC assumptions are violated.

As our main result, we establish in our Theorem 1 that the weight-robust treatment effect is characterized as the optimal value of a degenerate, constrained convex optimization problem, where degeneracy of the objective may lead to non-unique optimal solutions. Importantly, although the minimizers are not unique, the estimand itself (as the optimal value of the optimization problem) is uniquely identified and admits a clear interpretation as the most conservative treatment effect across all admissible post-treatment weights. This degeneracy, however, poses challenges for theoretical analysis: rather than attaining the standard parametric rate, the estimator converges more slowly, with the slower rate directly attributable to the structural degeneracy of the underlying optimization problem; see Theorem 3.

Despite being meaningful, statistical inference for the weight-robust treatment effect is challenging, because the DRoSC estimator may exhibit a non-normal limiting distribution, rendering conventional asymptotics unreliable. To address this, we develop a perturbation-based method for constructing valid confidence intervals. Our approach decomposes the DRoSC estimator’s uncertainty into two components: a regular part, handled with standard tools, and an irregular part driven by the constrained optimization geometry. To account for the latter, we generate a collection of perturbed optimization problems that preserve the geometry of the original formulation. This ensemble ensures that at least one perturbed instance nearly recovers the truth, enabling valid inference in the presence of non-regular asymptotics.

We evaluate the proposed method under diverse data-generating processes, including high correlations among control units and post-treatment weight shifts. As shown in Sections 7.2 and the Supplementary Material, the DRoSC estimator exhibits favorable finite-sample behavior and remains consistent in our settings. In regimes with non-regular asymptotics,

confidence intervals based on normal approximations suffer from undercoverage, whereas our perturbation-based intervals maintain nominal coverage; see Section 7.3 and the Supplementary Material. We further demonstrate the practical utility of DRoSC through a reanalysis of the Basque Country case study (Abadie and Gardeazabal, 2003).

To summarize, our main contributions are as follows:

- (1) We introduce the DRoSC method as a generalization of standard synthetic control. The DRoSC estimator targets a new causal effect—the *weight-robust treatment effect*—which remains interpretable and sign-consistent even when the true treatment effect is not point-identified.
- (2) We propose a perturbation-based inference procedure that delivers uniformly valid confidence intervals for the weight-robust treatment effect. The approach is of independent interest for non-regular uncertainty quantification arising from convex optimization problems with possibly non-unique solutions.

1.2 Other Related Works

Synthetic control. First introduced by Abadie and Gardeazabal (2003) and Abadie et al. (2010), the SC method has inspired a wide range of methodological developments; for a comprehensive review, see Abadie (2021) and references therein. Most existing studies impose identifiability conditions to recover unobserved potential outcomes, for example through linear prediction models (Li, 2020; Chernozhukov et al., 2021; Cattaneo et al., 2021; Shen et al., 2023; Chernozhukov et al., 2025), linear factor models (Xu, 2017; Shi et al., 2021; Ben-Michael et al., 2021), quantile functions (Gunsilius, 2023), or matrix completion approaches (Amjad et al., 2018; Bai and Ng, 2021; Athey et al., 2021). Most existing papers on the SC method focus on settings where the treatment effect is identifiable. In contrast, we study a practically relevant yet underexplored scenario in which identification fails due to weight shifts and high correlations among control units.

Sensitivity analysis and DRO. While some recent studies have incorporated sensitivity analyses to address violations of identification assumptions within the SC framework, these works primarily focus on different sources of weight misspecification. For instance, Zeitler et al. (2023) analyzed the bias induced by distributional shifts in latent causes under nonparametric models, yet their framework does not address the problem of high correlations among control units. Ferguson and Ross (2020) examined sensitivity to model misspecification, allowing true post-treatment weights outside the simplex, but did not formally account for statistical uncertainty. In contrast, we focus on a different regime of identification failure arising from weight shifts and high correlations among control units. Furthermore, we adopt a DRO-based framework that yields a single, interpretable causal estimand, in contrast to conventional sensitivity analysis approaches that characterize partial identification through sets of plausible values (e.g., Manski, 1990). Our work is the first to build an explicit connection between the DRO-based method and sensitivity analysis within the SC literature. Specifically, our identification theorem provides a geometric intuition of the weight-robust treatment effect being the most conservative treatment effect evaluated across all possible post-treatment weights within the uncertainty class. We provide a detailed comparison between sensitivity analysis and our DRO-based approach in Section 3.3.

Non-unique SC weight. The most relevant work addressing the issue of highly correlated controls or non-unique synthetic weight is Abadie and L’hour (2021), which discusses this

challenge in the context of multiple treated units and introduces a penalized SC method to promote uniqueness. Their approach constructs SC weights by penalizing discrepancies between the covariates of the treated and control units. However, this strategy is not applicable when only outcome variable is available (see, e.g., Doudchenko and Imbens, 2016; Amjad et al., 2018; Chernozhukov et al., 2021). More importantly, rather than mitigating non-uniqueness by adopting penalization (Abadie and L’hour, 2021), we address this issue with a distinct proposal through introducing an uncertainty class that comprehensively accounts for all possible synthetic weights.

Non-regular inference. Several studies have developed principled inference procedures within the SC framework, either by employing permutation-based approaches that avoid reliance on the asymptotic distribution of the estimated weights (e.g., Abadie et al., 2010; Hahn and Shi, 2017; Firpo and Possebom, 2018; Chernozhukov et al., 2021), or by relaxing the simplex constraint on the true weights so that the estimated weights are asymptotically normal (e.g., Shi et al., 2021; Shen et al., 2023). However, due to the simplex constraints, inference based on the estimator’s asymptotic distribution remains fundamentally challenging, even when the treatment effect is identifiable (Li, 2020; Cattaneo et al., 2021; Fry, 2024). In this paper, we focus on inference for the weight-robust treatment effect, which reduces to the true treatment effect in identifiable settings. Inference for this estimand remains difficult because boundary effects and instability induce non-regular behavior of our proposed DRoSC estimator. When estimators deviate from standard limiting laws, such as asymptotic normality, classical large-sample methods may fail to produce valid confidence intervals (Wasserman et al., 2020; Guo, 2023a; Xie and Wang, 2024; Kuchibhotla et al., 2024; Guo et al., 2025a,b), and both bootstrap and subsampling methods can break down under boundary constraints (Andrews, 1999). In Section 5.1, we show that the DRoSC estimator exhibits a non-regular limiting distribution due to the boundary constraint and instability. To address this, we introduce a perturbation-based inference procedure that targets the population version of the underlying optimization problem, as detailed in Section 5.2.

1.3 Organization and Notations

Our paper is organized as follows: Section 2 introduces our model setup, assumptions for the SC method, and associated identification challenges. Section 3 introduces the DRoSC method, detailing the identification of the weight-robust treatment effect, its interpretation, and its relationship to sensitivity analysis. Section 4 outlines the corresponding estimation procedure. Section 5 discusses inference challenges and introduces perturbation-based inference methods. Section 6 provides the convergence rate of the proposed estimator, and the coverage and precision properties of the proposed confidence interval. Sections 7 and 8 demonstrate the practical effectiveness of DRoSC through extensive simulations and an application to the Basque study (Abadie and Gardeazabal, 2003). Finally, Section 9 concludes and discusses some possible directions for further research.

We now introduce the notations used throughout the paper. For any vector $v \in \mathbb{R}^p$, the ℓ_q -norm of v is defined as $\|v\|_q = (\sum_{i=1}^p v_i^q)^{1/q}$ for $q \geq 0$ and $\|v\|_\infty$ denotes $\max_{1 \leq i \leq p} |v_i|$; $\mathbf{1}_p$ and $\mathbf{0}_p$ denote p -dimensional vectors where each coordinate takes the value 1 and 0 respectively; v_j denotes j -th element of v . For any $n \times p$ matrix M , $M^\top$ denotes the transpose of M ; $M_{i,j}$ denotes the entry of M in row i and column j ; $\|M\|_{\max}$ denotes $\max_{i,j} |M_{i,j}|$; $\mathbf{I}$ denotes the identity matrix with the corresponding dimension; for a symmetric matrix $M \in \mathbb{R}^{p \times p}$, $\lambda_{\min}(M)$ denotes the minimum eigenvalue of M . For a sequence x_n , we write $x_n \rightarrow x$ to

denote convergence, $x_n \xrightarrow{p} x$ to denote convergence in probability, and $x_n \xrightarrow{d} x$ to denote convergence in distribution. For positive sequences a_n and b_n , $a_n \lesssim b_n$ means that there exists $C > 0$ such that $a_n \leq Cb_n$ for all n ; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. *i.i.d* stands for “independent and identically distributed”. $\mathbb{I}(\cdot)$ denotes an indicator function. Throughout the paper, C denotes a generic positive constant, possibly varying from one occurrence to another.

2 Synthetic Control: Essential Assumptions and Challenges

2.1 Model Setup

We review the standard synthetic control setup with $N + 1$ units observed over a total of T time periods. The first unit receives control until time T_0 , after which it receives the treatment from time $T_0 + 1$ with $T_0 \leq T - 1$, while the remaining N units receive control for the whole T time periods. We refer to $t = 1, \dots, T_0$ as the pre-treatment period while $t = T_0 + 1, \dots, T$ as the post-treatment period.

We introduce the potential outcome notations with $Y_{j,t}^{(0)}$ and $Y_{j,t}^{(1)}$ representing the potential outcome of unit j at time t under control and treatment respectively. Since the first unit starts receiving the treatment from time $T_0 + 1$, the observed outcome for the first unit admits the following expression:

$$Y_{1,t} = \begin{cases} Y_{1,t}^{(0)} & \text{if } 1 \leq t \leq T_0, \\ Y_{1,t}^{(1)} & \text{if } T_0 + 1 \leq t \leq T. \end{cases} \quad (1)$$

The control units $2, \dots, N + 1$ do not receive the treatment over the entire period, and the observed outcomes satisfy $Y_{j,t} = Y_{j,t}^{(0)}$ for $2 \leq j \leq N + 1$ and $1 \leq t \leq T$. With the above notations, we define the average treatment effect on the treated (ATT) at time t (e.g., Shi et al., 2021; Park and Tchetgen Tchetgen, 2025) as

$$\tau_t = \mathbb{E} \left[Y_{1,t}^{(1)} - Y_{1,t}^{(0)} \right].$$

In the above definition, the expectation is taken with respect to the randomness of the potential outcomes $Y_{1,t}^{(1)}$ and $Y_{1,t}^{(0)}$ and we adopt the super-population framework (see, e.g., Imbens and Rubin, 2015, p. 99) throughout the paper. By the definition of τ_t , we write

$$Y_{1,t}^{(1)} - Y_{1,t}^{(0)} = \tau_t + v_t \quad \text{for } t = T_0 + 1, \dots, T, \quad (2)$$

where $\{v_t\}_{t=T_0+1}^T$ is a sequence of mean-zero random error terms arising from the randomness of the potential outcome differences $Y_{1,t}^{(1)} - Y_{1,t}^{(0)}$.

We further define the time-averaged ATT as:

$$\bar{\tau} = \frac{1}{T_1} \sum_{t=T_0+1}^T \tau_t \quad \text{with } T_1 = T - T_0. \quad (3)$$

In this paper, our primary focus is on inference for $\bar{\tau}$ (Arkhangelsky et al., 2021; Athey et al., 2021; Liu et al., 2024), or its conservative proxy when $\bar{\tau}$ is not identifiable. We allow for non-constant effects τ_t for $t = T_0 + 1, \dots, T$, thereby generalizing the constant-effect assumption adopted in Li (2020) and Shi et al. (2021). We shall remark that estimating the time-specific ATT τ_t is inherently a distinct and more challenging problem than our inference target $\bar{\tau}$ due to the availability of only a single treated unit at each post-treatment time point. Valid inference

for τ_t typically requires additional assumptions. For instance, some existing works assume a static treatment effect (i.e., $v_t = 0$) (e.g., Abadie et al., 2010), or impose a parametric model assumption on τ_t as a known function of t (Park and Tchetgen Tchetgen, 2025). Without such assumptions, researchers turn to constructing a prediction interval for the random quantity $Y_{1,t}^{(1)} - Y_{1,t}^{(0)}$ for each time t (e.g., Cattaneo et al., 2021). However, by focusing on the time-averaged ATT $\bar{\tau}$, it is more feasible to construct valid confidence intervals without the need to impose additional assumptions on τ_t or v_t , or turn to the construction of prediction intervals.

2.2 Essential Assumptions for the Synthetic Control Method

In this section, we discuss the essential assumptions ensuring that the SC method identifies $\bar{\tau}$ and shall emphasize in the following Section 2.3 on the practical scenarios that these identification conditions may fail. Throughout the paper, we facilitate the discussion by writing $X_t = (Y_{2,t}, \dots, Y_{N+1,t})^\top$ for $t = 1, \dots, T$, and consider the following models between the outcome of unit 1 and all other units (e.g. Chernozhukov et al., 2021; Shen et al., 2023),

$$Y_{1,t}^{(0)} = \begin{cases} X_t^\top \beta^{(0)} + u_t^{(0)} & \text{for } t = 1, \dots, T_0, \\ X_t^\top \beta^{(1)} + u_t^{(1)} & \text{for } t = T_0 + 1, \dots, T, \end{cases} \quad \text{with } \beta^{(0)}, \beta^{(1)} \in \Delta^N \quad (4)$$

where $\Delta^N = \{\beta : \beta_j \geq 0, \mathbf{1}_N^\top \beta = 1\}$ denotes the simplex in $\mathbb{R}^N$ and $\{u_t^{(0)}\}_{t=1}^{T_0}$ and $\{u_t^{(1)}\}_{t=T_0+1}^T$ are sequences of mean-zero error terms satisfying $\mathbb{E}[X_t u_t^{(0)}] = \mathbb{E}[X_t u_t^{(1)}] = \mathbf{0}_N$.

We now state the two critical assumptions for SC to identify $\bar{\tau}$ under the model (4).

(E1) $\beta^{(0)}$ is the unique minimizer of the constrained least squares for the pre-treatment period, that is,

$$\beta^{(0)} = \arg \min_{\beta \in \Delta^N} \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E} \left[\left(Y_{1,t} - X_t^\top \beta \right)^2 \right]. \quad (5)$$

(E2) There is no weight shift before and after the treatment, that is, $\beta^{(0)} = \beta^{(1)}$.

The main idea of the SC method is to identify $\beta^{(0)}$ as a solution of the optimization problem (5) and assume there is no weight shift and identify the effect $\bar{\tau}$ as $T_1^{-1} \sum_{t=T_0+1}^T \mathbb{E}[Y_{1,t} - X_t^\top \beta^{(0)}]$. This identification strategy motivates the following SC estimators of $\beta^{(1)}$ and $\bar{\tau}$,

$$\hat{\beta}^{\text{SC}} = \arg \min_{\beta \in \Delta^N} \frac{1}{T_0} \sum_{t=1}^{T_0} \left(Y_{1,t} - X_t^\top \beta \right)^2 \quad \text{and} \quad \hat{\tau}^{\text{SC}} = \frac{1}{T_1} \sum_{t=T_0+1}^T \left(Y_{1,t} - X_t^\top \hat{\beta}^{\text{SC}} \right). \quad (6)$$

2.3 Identification Challenges: Non-uniqueness and Weight Shift

In the following, we discuss how the critical identification conditions (E1) and (E2) may fail to hold in practice. First, the minimizer of the optimization problem (5) may not be unique when there are high correlations among control units' outcomes in the pre-treatment period, which is a common feature of data analyzed using the SC method. For example, in the Basque study introduced by Abadie and Gardeazabal (2003), who first proposed the SC method, Figure 1 shows that, during the pre-treatment period, the correlations between the SC-selected control units and all control units are predominantly close to one. Second, and equally important, the

relationship between the treated and control units may not remain stable in practice, as the treatment itself can affect the relationship between the potential outcomes in (4).

We present a semi-real data analysis using the Basque study dataset (Abadie and Gardeazabal, 2003; Abadie et al., 2011) and demonstrate that, when the identification conditions (E1) and (E2) fail, the SC estimator may become unreliable. Our analysis examines two key issues: (i) instability arising from the high correlations among control units, and (ii) bias in $\hat{\tau}^{SC}$ resulting from weight shifts. Detailed implementation steps for these experiments are provided in the Supplementary Material.

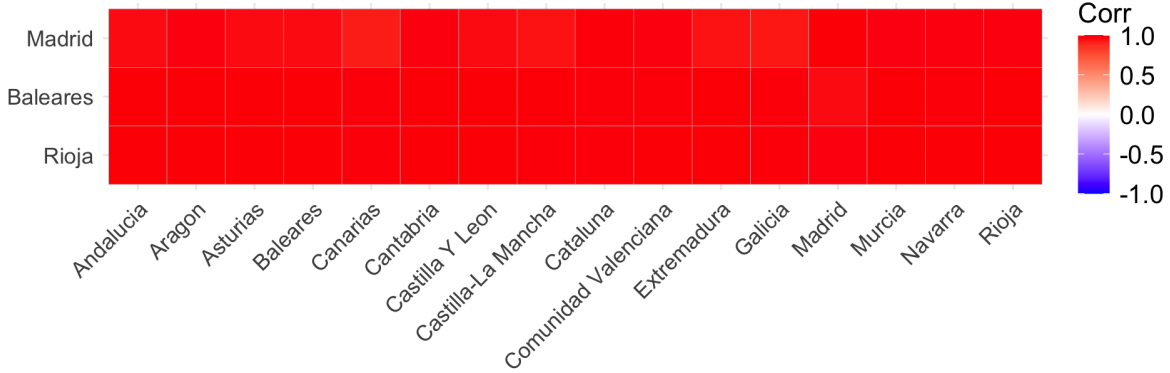


Figure 1: Correlation plot from the Basque study. The vertical axis denotes control units selected from SC, and the horizontal axis denotes all control units.

We first investigate the instability of $\hat{\beta}^{SC}$ by adding small random noise to the pre-treatment data and analyzing the resulting estimates. Specifically, we add i.i.d. normal random noise with standard error equal to c times that of the corresponding pre-treatment data with $c \in \{0.05, 0.1, 0.15\}$. Here, a larger value of c indicates a larger magnitude of added noise; see the Supplementary Material for the details of the data-generating process.

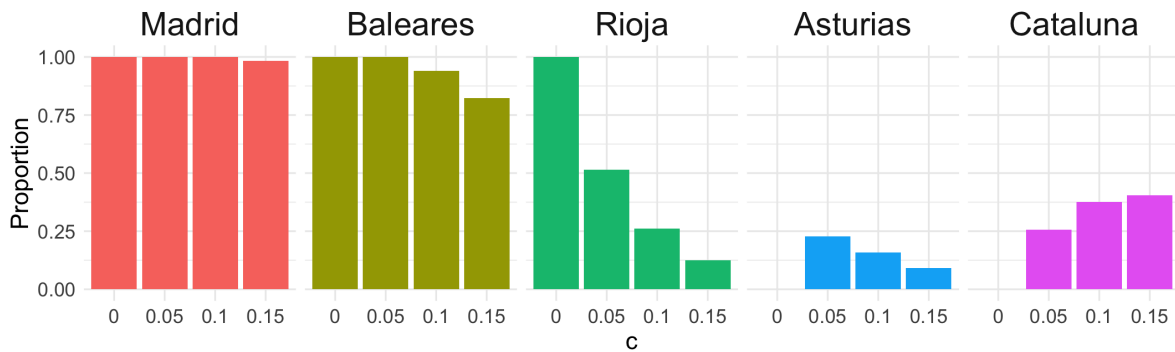


Figure 2: The proportion of the control units being selected by the SC method out of 1000 perturbed data sets. The variable c indicates the noise level applied to the dataset to generate the pre-treatment data.

For each level c , we report in the Figure 2 the proportion of times each control unit being selected by SC across 1000 simulated pre-treatment data. For the original data set (that is $c = 0$), the SC method selects Madrid, Baleares, and Rioja as the donors. As c increases, Madrid

remains consistently selected with a frequency close to 1 while the proportions of Balears and Rioja being selected decrease. Meanwhile, Asturias and Catalunya, whose outcomes are similar to those of Balears and Rioja, begin to be selected, with Catalunya’s proportion even increasing with c . This behavior happens due to the high correlations among control units, as shown in Figure 1, where multiple nearly equivalent weight combinations approximate the treated unit.

Next, we examine how weight shifts affect the SC estimator’s performance. We construct scenarios with weight shifts during the post-treatment period. Particularly, we generate the data with shifted weights on pairs of highly similar regions, Balears and Catalunya, and Rioja and Asturias. We generate semi-real data as follows: we set $\beta^{(0)}$ as the synthetic control weight estimator $\hat{\beta}^{\text{SC}}$ and set $\beta^{(1)}$ to have a weight shift from $\beta^{(0)}$, where, as specified on the left panel of Figure 3, the parameter κ controls the weight shift and takes values from 0.05 to 0.4. We generate 1000 semi-real data sets by adding independent small noise to the control units. We then generate the treated unit’s outcomes using model (4) with $\beta^{(0)}$ and $\beta^{(1)}$. We apply the SC method to each simulated dataset to estimate $\bar{\tau}$ via (6).

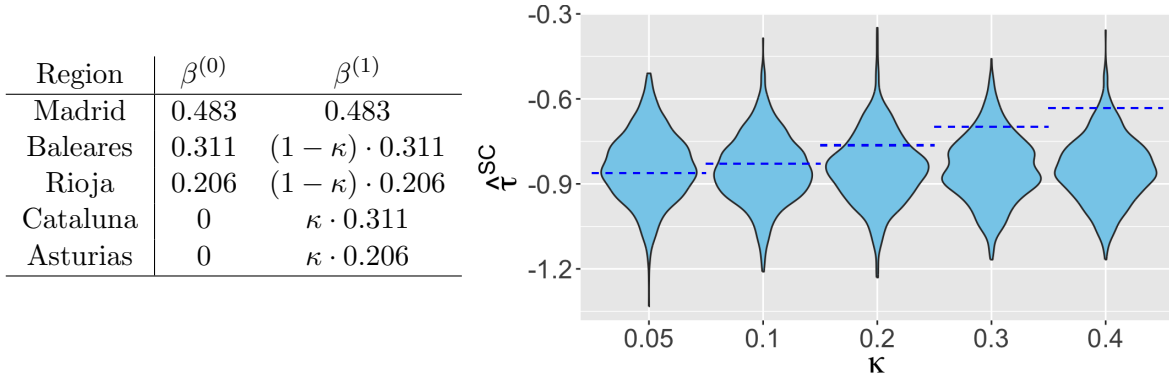


Figure 3: The table on the left displays the pre-treatment and post-treatment weights used to generate the model (4). The plot on the right shows violin plots of $\hat{\tau}^{\text{SC}}$, defined in (6), across 1000 simulations using newly generated data for each $\kappa \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$. The blue dashed line marks the true value of $\bar{\tau}$.

The right panel of Figure 3 presents the simulation results, illustrating how the performance of the SC estimator $\hat{\tau}^{\text{SC}}$ is affected by increasing degrees of weight shift. The horizontal axis represents the parameter κ , which controls the deviation between $\beta^{(0)}$ and $\beta^{(1)}$. As κ increases, $\beta^{(1)}$ diverges further from $\beta^{(0)}$. We use the blue dashed line to represent the true time-averaged ATT $\bar{\tau}$. Each violin plot at a given κ level displays the empirical distribution of the SC estimator $\hat{\tau}^{\text{SC}}$ across 1000 simulations. The center of each violin plot, representing the empirical average of $\hat{\tau}^{\text{SC}}$, remains relatively stable across increasing κ , indicating that the SC estimator fails to adjust for the changing post-treatment weights. As a result, the gap between the center of the violin plot and the blue dashed line (i.e., the bias of the SC estimator) widens as κ increases. This growing bias arises from the violation of condition (E2), which is necessary for identifying $\bar{\tau}$. Thus, Figure 3 highlights how unaccounted-for weight shifts lead to the standard SC method suffering from the bias and unreliable inference.

These findings highlight fundamental challenges in making reliable inferences using the SC method when its key identification conditions (E1) and (E2) fail: instability in the SC estimator $\hat{\tau}^{\text{SC}}$ due to high correlations among the control units and bias from post-treatment weight changes. These observations motivate us to consider a new causal estimand that provides

information for regimes where the SC identification conditions fail.

3 Distributionally Robust Synthetic Control

We have emphasized in Section 2.3 that the identification of the time-averaged ATT $\bar{\tau}$ becomes impossible when the key identification conditions fail. To address this, we introduce a new causal estimand, the weight-robust treatment effect, which aims to recover meaningful information about $\bar{\tau}$. When the critical identification conditions (E1) and (E2) hold, this new causal estimand is the same as $\bar{\tau}$. However, when these conditions fail, the new estimand serves as a conservative proxy for $\bar{\tau}$ in the sense that, when it is non-zero, $\bar{\tau}$ shares the same sign with this new causal estimand.

In Section 3.1, we define the weight-robust treatment effect through the lens of distributionally robust optimization. We present its identification in Section 3.2, and provide a thorough comparison to sensitivity analysis in Section 3.3.

3.1 New Causal Estimand via Distributionally Robust Optimization

In the following, we introduce a new causal estimand as a proxy of the time-averaged ATT $\bar{\tau}$ in (3). To motivate the definition, we express the time-averaged ATT $\bar{\tau}$ as the solution to an optimization problem and then generalize its definition through borrowing the strength of distributionally robust optimization.

We start with the optimization problem's objective function and then express $\bar{\tau}$ as a maximizer of the optimization problem in the following (7). For any given weight vector β , we define the following reward function $R_\beta(\tau)$ associated with the treatment effect τ ,

$$R_\beta(\tau) := \frac{1}{T_1} \sum_{t=T_0+1}^T \mathbb{E} \left[\left(Y_{1,t} - X_t^\top \beta \right)^2 - \left(Y_{1,t} - X_t^\top \beta - \tau \right)^2 \right].$$

For a given weight β and treatment effect τ , $R_\beta(\tau)$ compares the prediction error of using a null treatment effect and that of using a constant treatment effect τ . Hence, $R_\beta(\tau)$ represents the improvement in fit when a constant τ is incorporated in the post-treatment period. For a given β , our goal is to maximize $R_\beta(\tau)$, as a higher value indicates a greater reduction in prediction error under the assumption of a nonzero treatment effect.

When the oracle knowledge of $\beta^{(1)}$ is available, we write the time-averaged ATT $\bar{\tau}$ as the solution to the following optimization problem,

$$\bar{\tau} = \arg \max_{\tau \in \mathbb{R}} R_{\beta^{(1)}}(\tau). \quad (7)$$

Note that, for a given $\beta \in \Delta^N$, the maximizer of $R_\beta(\tau)$ is given by

$$\tau(\beta) = \mu_Y - \mu^\top \beta, \quad \text{with} \quad \mu_Y = \frac{1}{T_1} \sum_{t=T_0+1}^T \mathbb{E}[Y_{1,t}] \quad \text{and} \quad \mu = \frac{1}{T_1} \sum_{t=T_0+1}^T \mathbb{E}[X_t]. \quad (8)$$

It follows from the definition of (3) and the model (4) that $\bar{\tau} = \tau(\beta^{(1)})$.

Despite we write $\bar{\tau}$ as the solution to the optimization problem in (7), the identification challenge of $\bar{\tau}$ persists since identification of $\beta^{(1)}$ relies on stringent assumptions that may not hold in practice. To address this challenge, rather than attempting to identify the true $\beta^{(1)}$

by imposing conditions (E1) and (E2), we introduce a new estimand based on distributionally robust optimization (DRO) (see, e.g., Ben-Tal et al., 2009; Duchi and Namkoong, 2021, and references therein). In particular, we define the following uncertainty class: for $\lambda \geq 0$,

$$\Omega(\lambda) := \left\{ \beta \in \Delta^N : \left\| \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E} \left[X_t (Y_{1,t} - X_t^\top \beta) \right] \right\|_\infty \leq \lambda \right\}. \quad (9)$$

The uncertainty class $\Omega(\lambda)$ consists of all weight vectors β for which the time-averaged covariance between the control units' outcomes X_t and the residuals $Y_{1,t} - X_t^\top \beta$ over the pre-treatment period is uniformly small across all N control units. The parameter $\lambda \geq 0$ controls the degree of allowable deviation: when $\lambda = 0$, the set reduces to all weights that exactly balance moments in expectation, including the true pre-treatment weight $\beta^{(0)}$. As λ increases, the class $\Omega(\lambda)$ permits larger moment imbalances, thereby encompassing a wider range of post-treatment weights, with the hope of containing the true post-treatment weight $\beta^{(1)}$ for a properly chosen λ . In this sense, we treat all weight vectors with imbalance level up to λ as potential candidates for $\beta^{(1)}$. We shall emphasize that the choice of λ reflects the user's belief about the extent of weight shift. In practice, one may adopt a strategy similar to sensitivity analysis by examining results across a range of λ values (see, e.g., Rosenbaum, 2002).

Since $\beta^{(1)}$ can be any element of Ω , we enumerate all possible weights belonging to Ω and define the worst-case reward of the treatment effect τ over Ω as

$$\min_{\beta \in \Omega} R_\beta(\tau). \quad (10)$$

By taking the minimum over all possible post-treatment weights, this formulation captures the worst-case scenario for the treatment effect, considering the most challenging configurations of the weights within Ω . Similar to (7), we define the weight-robust treatment effect as the optimizer of the worst-case reward (10):

$$\tau^*(\Omega) := \arg \max_{\tau \in \mathbb{R}} \left[\min_{\beta \in \Omega} R_\beta(\tau) \right]. \quad (11)$$

When the identification conditions (E1) and (E2) hold such that $\bar{\tau}$ is identified by the standard SC method, this new estimand $\tau^*(\Omega)$ is reduced to $\bar{\tau}$ by using Ω with $\lambda = 0$. However, even when $\bar{\tau}$ is not identifiable, $\tau^*(\Omega)$ is still identifiable and serves as a conservative proxy of $\bar{\tau}$ as established in the following Corollary 1. From a game-theoretic perspective (Blackwell and Girshick, 1979), this framework can be interpreted as a two-player game: nature adversarially selects the worst-case post-treatment weight from Ω , while the decision-maker chooses τ to maximize the resulting reward. The resulting $\tau^*(\Omega)$ thus represents a treatment effect for adversarially chosen post-treatment weights.

3.2 Identification and Interpretation

In this subsection, we present the identification theorem of $\tau^*(\Omega)$ defined in (11), which enables us to design a data-dependent estimator of $\tau^*(\Omega)$ in Section 4.

Theorem 1. *$\tau^*(\Omega)$ defined in (11) is uniquely identified as*

$$\tau^*(\Omega) = \mu_Y - \mu^\top \beta^*(\Omega), \quad \text{where} \quad \beta^*(\Omega) = \arg \min_{\beta \in \Omega} \left[\mu_Y - \mu^\top \beta \right]^2. \quad (12)$$

Theorem 1 provides a method for explicitly computing $\tau^*(\Omega)$ by first identifying the adversarial weight $\beta^*(\Omega)$ through solving a quadratic optimization problem and then applying this synthetic weight to identify $\tau^*(\Omega)$ as $\mu_Y - \mu^\top \beta^*(\Omega)$. Intuitively, $\beta^*(\Omega)$ is the post-treatment weight within Ω that makes the time-averaged ATT as close to zero as possible, highlighting that $\tau^*(\Omega)$ is the most conservative treatment effect. We note that the identification of τ^* in Theorem 1 does not rely on the key assumptions of the SC method. When there is no ambiguity, we denote $\tau^*(\Omega)$ and $\beta^*(\Omega)$ as τ^* and β^* , respectively.

While Theorem 1 provides a straightforward characterization of β^* and hence τ^* , the quadratic program in (12) is a degenerate convex optimization problem because $\mu\mu^\top$ is rank one. Even though the optimal value τ^* is uniquely defined, this degenerate optimization problem implies that the optimizer β^* may not be unique. The degenerate quadratic optimization problem in (12) complicates estimation, inference, and the associated theoretical analysis, since the optimal solution β^* directly linked to the optimization problem may not be uniquely defined. Particularly, with such a degenerate quadratic optimization problem, we are only able to show that our proposed estimator attains a slow convergence rate instead of the parametric rate; see Theorem 3 in Section 6.

Building on Theorem 1, we introduce the following theorem that provides an interpretation of our proposed new causal estimand τ^* .

Theorem 2. *For τ^* defined in (11), we attain the following equivalent expression:*

$$\tau^* = \begin{cases} \min_{\beta \in \Omega} \tau(\beta) & \text{if } \tau(\beta) > 0 \text{ for all } \beta \in \Omega, \\ \max_{\beta \in \Omega} \tau(\beta) & \text{if } \tau(\beta) < 0 \text{ for all } \beta \in \Omega, \\ 0 & \text{if there exists } \beta \in \Omega \text{ such that } \tau(\beta) = 0. \end{cases}$$

where $\tau(\beta)$ is defined in (8).

The above theorem establishes that τ^* corresponds to the point within the range of $\{\tau(\beta)\}_{\beta \in \Omega}$ that is closest to the origin. Intuitively, this means that we consider all possible time-averaged ATTs with synthetic post-treatment weights in Ω and then the new causal estimand τ^* represents the most conservative time-averaged ATT. We illustrate this geometric interpretation in Figure 4.

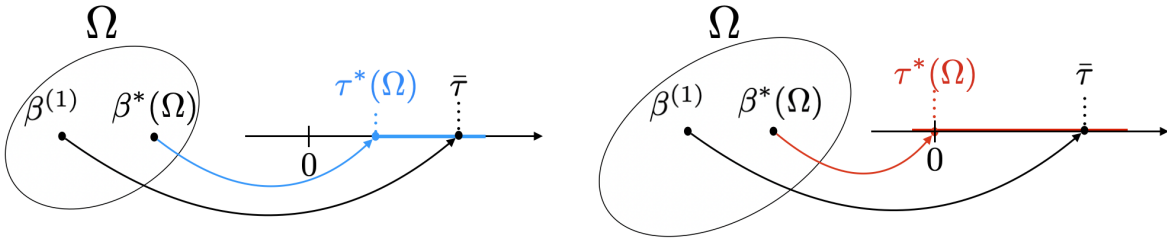


Figure 4: The ellipsoids in the both panels denote the uncertainty class Ω . The blue and red lines on the axis denote the range of $\{\tau(\beta)\}_{\beta \in \Omega}$ and $\tau^* = \tau^*(\Omega)$ denotes the nearest point to zero among all values in the interval.

Theorem 2 leads to the following corollary, highlighting the connection between $\tau^*(\Omega)$ and the time-averaged ATT $\bar{\tau}$ when $\beta^{(1)} \in \Omega$.

Corollary 1. *If $\beta^{(1)} \in \Omega$, then $\tau^*(\Omega)$ does not have an opposite sign to $\bar{\tau}$ and $|\tau^*(\Omega)| \leq |\bar{\tau}|$.*

Corollary 1 shows that $\tau^*(\Omega)$ serves as a conservative proxy for $\bar{\tau}$ as Ω is large enough to include $\beta^{(1)}$. When Conditions (E1) and (E2) hold, ensuring that $\bar{\tau}$ is identifiable, τ^* coincides with $\bar{\tau}$ by setting $\lambda = 0$. However, when the exact recovery of $\bar{\tau}$ is infeasible due to the failure of (E1) or (E2), τ^* provides a conservative approximation: its magnitude is bounded from above by that of $\bar{\tau}$ and the sign of τ^* will not be opposite to that of $\bar{\tau}$. When $\tau^*(\Omega)$ is non-zero, $\tau^*(\Omega)$ shares the same sign as $\bar{\tau}$. Figure 5 illustrates how τ^* operates in both identifiable and non-identifiable settings.

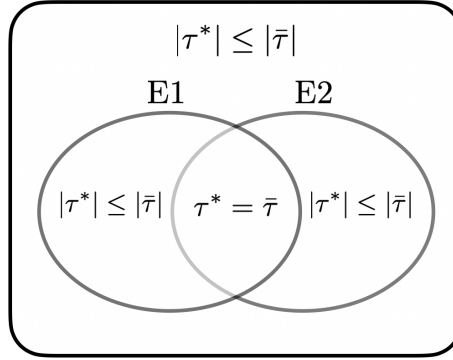


Figure 5: Relationship between τ^* and $\bar{\tau}$ when the uncertainty class Ω contains $\beta^{(1)}$. The circles labeled E1 and E2 represent the settings where Conditions (E1) and (E2) are satisfied, respectively. The rectangle indicates the general case, including scenarios where these assumptions do not hold.

The following proposition quantifies the difference between τ^* and $\bar{\tau}$.

Proposition 1. *If $\lambda_{\min}(\Sigma) > 0$ and $\beta^{(1)} \in \Omega$, then*

$$|\tau^* - \bar{\tau}| \leq 2[\lambda_{\min}(\Sigma)]^{-1} \|\mu\|_1 \sqrt{N} \lambda.$$

When (E1) and (E2) hold, we can take $\lambda = 0$, in which case $\tau^* = \bar{\tau}$. More generally, τ^* recovers $\bar{\tau}$ closely when $\lambda \rightarrow 0$ sufficiently fast such that the right hand side of the above inequality converges to zero. For example, if $\{X_t\}_{t=T_0+1}^T$ is stationary so that $\|\mu\|_1$ is fixed, and if $\lambda_{\min}(\Sigma)$ is bounded away from zero with fixed N , then $\tau^* \rightarrow \bar{\tau}$ as $\lambda \rightarrow 0$.

3.3 Comparison to Sensitivity Analysis

Theorem 2 and Corollary 1 enable us to establish a connection between our framework and sensitivity analysis. Using the sensitivity parameter λ , we can define a sensitivity model for the post-treatment weight, denoted as Ω , given in (9). Assuming the true weight $\beta^{(1)}$ belongs to Ω as required in Corollary 1, the following sensitivity interval contains $\bar{\tau}$, enabling its partial identification (Manski, 1990):

$$\left[\min_{\beta \in \Omega} \tau(\beta), \max_{\beta \in \Omega} \tau(\beta) \right]. \quad (13)$$

While sensitivity analysis aims to identify an interval for $\bar{\tau}$, our approach introduces τ^* as a point estimand, the most conservative time-averaged ATT within the interval (13). A nonzero τ^* implies that the interval (13) excludes zero, thereby enabling identification of $\text{sgn}(\bar{\tau})$, a goal shared by both frameworks, as established in Corollary 1. Despite this commonality,

the two approaches differ fundamentally in philosophy: sensitivity analysis is grounded in partial identification, whereas our method is rooted in distributional robustness. Building on this perspective, τ^* is a causal estimand that summarizes the interval (13) and provides meaningful information, particularly when the sign of the time-averaged ATT is of primary interest.

4 Estimation Procedure

In the following, we devise a data-dependent estimator of β^* and τ^* utilizing the population identification established in Theorem 1. We begin by constructing an estimator for the uncertainty class Ω . With $\Sigma = T_0^{-1} \sum_{t=1}^{T_0} \mathbb{E}[X_t X_t^\top]$ and $\gamma = T_0^{-1} \sum_{t=1}^{T_0} \mathbb{E}[X_t Y_{1,t}]$, we express the uncertainty class Ω in (9) as

$$\Omega = \{\beta \in \Delta^N : \|\gamma - \Sigma\beta\|_\infty \leq \lambda\}. \quad (14)$$

We construct the following data-dependent estimator of Ω

$$\hat{\Omega}(\lambda) = \left\{ \beta \in \Delta^N : \|\hat{\gamma} - \hat{\Sigma}\beta\|_\infty \leq \lambda + \rho \right\}, \quad \text{with} \quad \hat{\Sigma} = \frac{1}{T_0} \sum_{t=1}^{T_0} X_t X_t^\top, \quad \hat{\gamma} = \frac{1}{T_0} \sum_{t=1}^{T_0} X_t Y_{1,t}, \quad (15)$$

where ρ is a tuning parameter of order $[\log(\max\{T_0, N\})/T_0]^{1/2}$ that is introduced to account for the estimation errors $\hat{\Sigma} - \Sigma$ and $\hat{\gamma} - \gamma$. We provide a detailed discussion on how to choose the tuning parameter ρ in a data-dependent way towards the end of this section. For simplicity, when there is no ambiguity, we denote $\hat{\Omega}(\lambda)$ as $\hat{\Omega}$.

Building on the identification in Theorem 1, we estimate β^* by solving the following optimization problem:

$$\hat{\beta}(\hat{\Omega}) := \arg \min_{\beta \in \hat{\Omega}} \left[\hat{\mu}_Y - \hat{\mu}^\top \beta \right]^2, \quad (16)$$

where we estimate Ω by $\hat{\Omega}$ defined in (15) and estimate μ_Y and μ in $\tau(\beta)$, as defined in (8), by

$$\hat{\mu}_Y = \frac{1}{T_1} \sum_{t=T_0+1}^T Y_{1,t}, \quad \hat{\mu} = \frac{1}{T_1} \sum_{t=T_0+1}^T X_t. \quad (17)$$

We further estimate τ^* by

$$\hat{\tau}(\hat{\Omega}) = \hat{\mu}_Y - \hat{\mu}^\top \hat{\beta}(\hat{\Omega}). \quad (18)$$

For convenience, we respectively denote $\hat{\beta}(\hat{\Omega})$ and $\hat{\tau}(\hat{\Omega})$ as $\hat{\beta}$ and $\hat{\tau}$ if there is no confusion. Since we are borrowing the idea from distributional robustness, we shall refer to our estimation procedure as Distributionally Robust Synthetic Control (DRoSC), with details provided in Algorithm 1.

Algorithm 1 Distributionally Robust Synthetic Control (DRoSC)

Input: Pre-treatment data $\{Y_{1,t}, X_t\}_{t=1}^{T_0}$; Post-treatment data $\{Y_{1,t}, X_t\}_{t=T_0+1}^T$; Weight shift parameter $\lambda \geq 0$; Tuning parameter $\rho \geq 0$.

Output: Point estimator $\hat{\beta}$ of β^* ; Point estimator $\hat{\tau}$ of τ^* .

1: Construct the uncertainty class $\hat{\Omega}$ as in (15);

2: Construct $\hat{\mu}_Y$ and $\hat{\mu}$ as in (17);

3: Construct $\hat{\beta}$ as in (16) and $\hat{\tau}$ as in (18).

▷ DRoSC estimator

Tuning Parameter Selection. For the estimation of Ω in Section 4, we substitute γ and Σ in (14) with $\hat{\gamma}$ and $\hat{\Sigma}$. As a result of these substitutions, we introduce an additional tuning parameter ρ to account for the additional estimation error. For the i.i.d. data, the theoretical result suggests choosing ρ with a data-dependent way as

$$\rho = C \left[\hat{\sigma} \cdot \max_{2 \leq j \leq N+1} \left(\frac{1}{T_0} \sum_{t=1}^{T_0} Y_{j,t}^2 \right)^{\frac{1}{2}} + \lambda \right] \frac{[\log(\max\{T_0, N\})]^{1/2}}{\sqrt{T_0}} \quad (19)$$

where $\hat{\sigma}^2 = (T_0 - 1)^{-1} \sum_{t=1}^{T_0} (Y_{1,t} - X_t^\top \hat{\beta}^{\text{SC}})^2$ for some positive constant $C > 0$. We provide the theoretical justification of this choice of ρ in the Supplementary Material. However, we still need to identify the exact constant C in (19). To resolve this, we first start with the small value, such as $C = 0.01$. With the initial value specified in (19) with $C = 0.01$, however, it is possible that no feasible solution exists for (16) with the estimated uncertainty class $\hat{\Omega}$ in (15). To address this, we incrementally increase C by a factor of 1.25 until a feasible solution to (16) is obtained. We apply this iterative algorithm to identify the smallest value of ρ for which the optimization problem (16) admits a feasible solution.

While we specify ρ to be of order $[\log(\max\{T_0, N\})]^{1/2}/\sqrt{T_0}$ in (19) or the i.i.d. regime, our theoretical justification in Assumption 1 in Section 6 allows for a more general form: $\rho = [\log(\max\{T_0, N\})]^a/\sqrt{T_0}$, where the exponent $a \geq 1/2$ reflects the temporal dependence structure of the pre-treatment data. Larger values of a accommodate stronger temporal dependence. In the case where the pre-treatment data are i.i.d., our default specification with $a = 1/2$ suffices. However, when serial dependence is present, using $a > 1/2$ is necessary to ensure valid estimation. In practice, since we adopt our aforementioned procedure of choosing the constant in ρ , the choice of a does not strongly affect the final value of ρ . Large-scale numerical studies confirm that varying a does not affect the ability to conduct reliable estimation; see the Supplementary Material.

5 DRoSC Inference: Perturbation-based Methods

We now turn to the statistical inference for our new causal estimand τ^* . We start with demonstrating its inference challenge in Section 5.1 and devise a novel perturbation-based inference in Section 5.2.

5.1 Inference Challenge: Non-regularity and Instability

The inference challenge arises because the estimator $\hat{\tau}$, defined in (18), may not follow a standard limiting distribution. Specifically, the estimation error $\hat{\tau} - \tau^*$ can be decomposed as

follows:

$$\hat{\tau} - \tau^* = \hat{\mu}_Y - \mu_Y - \left(\hat{\mu}^\top \hat{\beta} - \mu^\top \beta^* \right), \quad (20)$$

where the decomposition decouples the randomness into two components: one for estimating μ_Y and the other for $\mu^\top \beta^*$. Although we can justify the asymptotic normality of $\hat{\mu}_Y - \mu_Y$, the term $\hat{\mu}^\top \hat{\beta} - \mu^\top \beta^*$ may exhibit a non-regular distribution due to the boundary constraint on β^* and the high correlations among the control units.

We illustrate the inference challenges associated with $\hat{\tau} - \tau$ in Figure 6 and shall explain its reasoning right after presenting the results. In Figure 6, we plot histograms of $\hat{\tau}$ and $\hat{\mu}^\top \hat{\beta}$ based on 500 simulations with $T_0 = 25$ and $T_1 = 25$ and detail the simulation settings in Section 7.1. The leftmost panel corresponds to the setting (S2) with $\tau^* \approx -0.6$, whose details are provided in Section 7.1. It shows a favorable setting where the limiting distribution of the DRoSC estimator $\hat{\tau}$ is nearly normal, suggesting that confidence intervals (CIs) based on normality achieve the desired coverage. In contrast, the middle and rightmost panels, which correspond to the settings (S3) with $\tau^* \approx 0.84$ and (S2) with $\tau^* \approx 0.05$ respectively, show settings where the distribution of $\hat{\tau}$ deviates from normality. The non-regular limiting distribution leads to undercoverage of CIs constructed under normality assumptions.

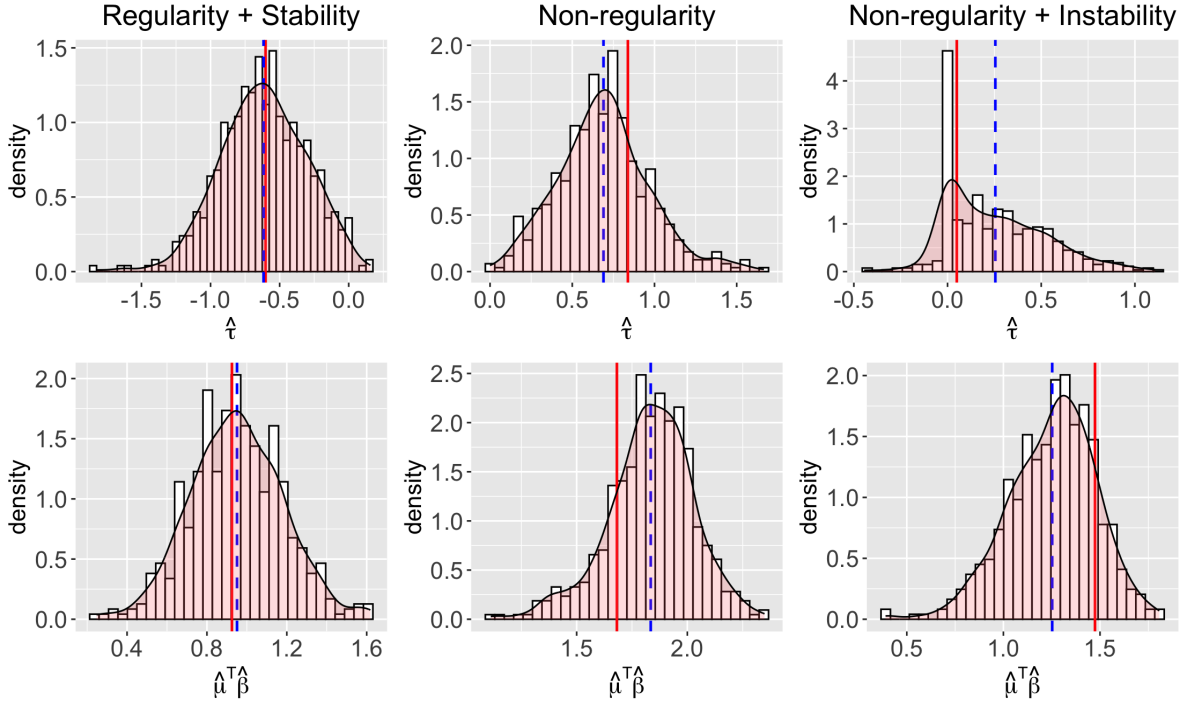


Figure 6: Histograms of $\hat{\tau}$ and $\hat{\mu}^\top \hat{\beta}$ based on 500 simulations from Section 7.1. The figures, from left to right, correspond to the settings (S2) with $\bar{\tau} = -1$ ($\tau^* \approx -0.6$), (S3) with $\bar{\tau} = 0.9$ ($\tau^* \approx 0.84$), and (S2) with $\bar{\tau} = 0.2$ ($\tau^* \approx 0.05$) with $T_0 = 25$ and $T_1 = 25$. In the top panel, the red solid line indicates τ^* , while in the bottom panel it indicates $\mu^\top \beta^*$. The blue dashed lines in both panels represent the sample averages across the 500 simulations.

We now explain that the inference challenges for $\hat{\tau} - \tau$, illustrated in Figure 6, arise primarily from the difficulty of quantifying the uncertainty in $\hat{\mu}^\top \hat{\beta} - \mu^\top \beta^*$. The first challenge, termed as non-regularity, results from the boundary constraint on β^* . Such a boundary constraint

leads to $\hat{\mu}^\top \hat{\beta} - \mu^\top \beta^*$ following a mixture distribution and boundary constraints are known to lead to non-regular inference (see, e.g., Self and Liang, 1987; Andrews, 1999; Drton, 2009; Guo, 2023b). Consequently, conventional inference methods relying on asymptotic normality or the bootstrap fail to provide valid results (e.g., Andrews, 2000). In addition to the non-regularity induced by the boundary constraint on β^* , the distribution of $\hat{\mu}^\top \hat{\beta}$ can be unstable when the control units are highly correlated. In such cases, the constraint set Ω can be nearly flat along certain directions, so that small estimation errors in γ or Σ can lead to substantial perturbations in the estimated uncertainty class $\hat{\Omega}$ in (15). More severely, if β^* lies near the boundary of Ω and the control units are highly correlated, even slight estimation errors in Ω can cause $\hat{\beta}$ to cross between the interior and exterior of the boundary, thereby inducing non-regularity and instability simultaneously.

5.2 Perturbation-based Inference

To address the challenge outlined in Section 5.1, we propose in the following a novel perturbation method. We start with presenting the main intuition of our proposal. To illustrate the main idea, we recall the following identification result established in Theorem 1,

$$\beta^* = \arg \min_{\beta \in \Omega} \left(\mu_Y - \mu^\top \beta \right)^2 \quad \text{with} \quad \Omega = \{ \beta \in \Delta^N : \|\gamma - \Sigma \beta\|_\infty \leq \lambda \}. \quad (21)$$

The data-driven estimator $\hat{\beta}$ presented in (16) is to replace $\{\Sigma, \gamma, \mu_Y, \mu\}$ in (21) with their sample analogs $\{\hat{\Sigma}, \hat{\gamma}, \hat{\mu}_Y, \hat{\mu}\}$. Our main idea is to create a collection of perturbed optimization problems, where the perturbation is added to the sample-based optimization problem in (16). Our main objective is to ensure that one of these perturbed optimization problems almost recovers the population optimization problem in (21).

We generate the perturbed optimization problems by adding perturbations to the sample average $\{\hat{\Sigma}, \hat{\gamma}, \hat{\mu}_Y, \hat{\mu}\}$. After constructing M perturbed quantities $\{\hat{\Sigma}^{[m]}, \hat{\gamma}^{[m]}, \hat{\mu}_Y^{[m]}, \hat{\mu}^{[m]}\}_{m=1}^M$, we use each set of perturbed quantities to define a corresponding perturbed optimization problem, and solve for the corresponding weight vector $\hat{\beta}^{[m]}$ as follows:

$$\hat{\beta}^{[m]} = \arg \min_{\beta \in \hat{\Omega}^{[m]}(\lambda)} \left[\hat{\mu}_Y^{[m]} - (\hat{\mu}^{[m]})^\top \beta \right]^2, \quad (22)$$

where the perturbed uncertainty class $\hat{\Omega}^{[m]}(\lambda)$, defined in the following (26), is constructed with $\hat{\gamma}^{[m]}$ and $\hat{\Sigma}^{[m]}$. We show that there exists m^* such that the perturbed optimization problem in (22) with $m = m^*$ nearly recovers the population optimization problem in (21) and hence $(\hat{\mu}^{[m^*]})^\top \hat{\beta}^{[m^*]}$ is nearly the same as $\mu^\top \beta^*$; see the theorem in the Supplementary Material. The remaining uncertainty lies primarily in estimating μ_Y , which can be addressed using standard inference methods, such as those based on asymptotic normality.

We now provide full details of our proposal. We assume that the estimators $\hat{\Sigma}$ and $\hat{\gamma}$ defined in (15), and $\hat{\mu}_Y$ and $\hat{\mu}$ defined in (17), satisfy the following asymptotic approximations:

$$\text{vecl}(\hat{\Sigma} - \Sigma) \stackrel{d}{\approx} \mathcal{N}(0, \hat{\mathbf{V}}_\Sigma), \quad \hat{\gamma} - \gamma \stackrel{d}{\approx} \mathcal{N}(0, \hat{\mathbf{V}}_\gamma), \quad \hat{\mu}_Y - \mu_Y \stackrel{d}{\approx} \mathcal{N}(0, \hat{\mathbf{V}}_Y), \quad \hat{\mu} - \mu \stackrel{d}{\approx} \mathcal{N}(0, \hat{\mathbf{V}}_\mu),$$

where $\text{vecl}(\hat{\Sigma} - \Sigma)$ is the vector formed by stacking the columns of the lower triangle part of $\hat{\Sigma} - \Sigma$, and $\stackrel{d}{\approx}$ denotes approximate equality in distribution and $\hat{\mathbf{V}}_\Sigma$, $\hat{\mathbf{V}}_\gamma$, $\hat{\mathbf{V}}_Y$, and $\hat{\mathbf{V}}_\mu$ denote

estimated covariance matrices defined as

$$\begin{aligned}
\widehat{\mathbf{V}}_\Sigma &= \frac{1}{T_0(T_0 - 1)} \sum_{t=1}^{T_0} \left(\text{vecl}(X_t X_t^\top) - \text{vecl}(\widehat{\Sigma}) \right) \left(\text{vecl}(X_t X_t^\top) - \text{vecl}(\widehat{\Sigma}) \right)^\top, \\
\widehat{\mathbf{V}}_\gamma &= \frac{1}{T_0(T_0 - 1)} \sum_{t=1}^{T_0} (X_t Y_{1,t} - \widehat{\gamma}) (X_t Y_{1,t} - \widehat{\gamma})^\top, \\
\widehat{\mathbf{V}}_Y &= \frac{1}{T_1(T_1 - 1)} \sum_{t=T_0+1}^T (Y_{1,t} - \widehat{\mu}_Y)^2, \quad \widehat{\mathbf{V}}_\mu = \frac{1}{T_1(T_1 - 1)} \sum_{t=T_0+1}^T (X_t - \widehat{\mu}) (X_t - \widehat{\mu})^\top.
\end{aligned} \tag{23}$$

We detail in the following our two-step proposal: perturbation and aggregation.

Step 1: Perturbation. We begin by generating perturbed quantities related to the uncertainty class Ω in (14) and objective function $\tau(\beta)$ in (8). Specifically, conditioning on the observed data, we generate i.i.d. samples $\{\widehat{\Sigma}^{[m]}\}_{m=1}^M$ and $\{\widehat{\gamma}^{[m]}\}_{m=1}^M$, following

$$\text{vecl}(\widehat{\Sigma}^{[m]}) \sim \mathcal{N} \left(\text{vecl}(\widehat{\Sigma}), \widehat{\mathbf{V}}_\Sigma + \|\widehat{\mathbf{V}}_\Sigma\|_{\max} \mathbf{I} \right), \quad \widehat{\gamma}^{[m]} \sim \mathcal{N} \left(\widehat{\gamma}, \widehat{\mathbf{V}}_\gamma + \|\widehat{\mathbf{V}}_\gamma\|_{\max} \mathbf{I} \right). \tag{24}$$

To ensure the symmetry of $\widehat{\Sigma}^{[m]}$, we impute the upper triangle part of each perturbed matrix $\widehat{\Sigma}^{[m]}$ by setting $\widehat{\Sigma}_{k,l}^{[m]} = \widehat{\Sigma}_{l,k}^{[m]}$ for $1 \leq l < k \leq N$. In addition, conditioning on the observed data, we generate i.i.d. samples $\{\widehat{\mu}^{[m]}\}_{m=1}^M$ and $\{\widehat{\mu}_Y^{[m]}\}_{m=1}^M$, which are related to the objective function $\tau(\beta)$, following

$$\widehat{\mu}_Y^{[m]} \sim \mathcal{N} \left(\widehat{\mu}_Y, \widehat{\mathbf{V}}_Y \right), \quad \widehat{\mu}^{[m]} \sim \mathcal{N} \left(\widehat{\mu}, \widehat{\mathbf{V}}_\mu + \|\widehat{\mathbf{V}}_\mu\|_{\max} \mathbf{I} \right). \tag{25}$$

We add a diagonal matrix to the corresponding covariance matrix in the above generating process such that the covariance matrix is positive definite. Specifically, we slightly enlarge the covariance matrices $\widehat{\mathbf{V}}_\Sigma$, $\widehat{\mathbf{V}}_\gamma$, and $\widehat{\mathbf{V}}_\mu$ to $\widehat{\mathbf{V}}_\Sigma + \|\widehat{\mathbf{V}}_\Sigma\|_{\max} \mathbf{I}$, $\widehat{\mathbf{V}}_\gamma + \|\widehat{\mathbf{V}}_\gamma\|_{\max} \mathbf{I}$, and $\widehat{\mathbf{V}}_\mu + \|\widehat{\mathbf{V}}_\mu\|_{\max} \mathbf{I}$, respectively. This adjustment mitigates numerical instability arising from near-singular covariance matrices, especially when N is relatively large compared to T_0 or T_1 ¹.

Throughout the paper, we use $p = 1 + N(N + 5)/2$ to represent the total dimensionality of the quantities $\text{vecl}(\Sigma)$, γ , μ_Y , and μ that are used in the population optimization problem in (21). For $1 \leq m \leq M$, we substitute $\widehat{\Sigma}$ and $\widehat{\gamma}$ in (15) with the perturbed ones $\widehat{\Sigma}^{[m]}$ and $\widehat{\gamma}^{[m]}$ and construct the perturbed uncertainty class $\widehat{\Omega}^{[m]}(\lambda)$ as

$$\widehat{\Omega}^{[m]}(\lambda) = \left\{ \beta \in \Delta^N : \|\widehat{\gamma}^{[m]} - \widehat{\Sigma}^{[m]} \beta\|_\infty \leq \lambda + \rho_M \right\}, \tag{26}$$

where $\rho_M \asymp [\log(\min\{T_0, T_1\})/M]^{1/p} / \sqrt{T_0}$ is a tuning parameter. The theorem in the Supplementary Material implies that there exists a perturbation index $m = m^*$ for which $\widehat{\Sigma}^{[m]}$ and $\widehat{\gamma}^{[m]}$ closely recover Σ and γ up to an error of order ρ_M ; with this $m = m^*$, the resulting perturbed uncertainty class $\widehat{\Omega}^{[m]}(\lambda)$ recovers the population uncertainty class Ω . We defer the discussion of selecting the tuning parameter ρ_M after providing the full procedure of our proposed method. When there is no confusion, we denote $\widehat{\Omega}^{[m]}(\lambda)$ as $\widehat{\Omega}^{[m]}$.

Using perturbed quantities $\widehat{\mu}_Y^{[m]}$ and $\widehat{\mu}^{[m]}$ for the objective function, we construct the perturbed weight vectors $\{\widehat{\beta}^{[m]}\}_{m \in \mathbb{M}}$ as the optimizer of the minimization problem (22). We show

¹While the main paper focuses on the regime where T_0 and T_1 are large relative to a fixed N , in practice, it is possible for N to exceed either T_0 or T_1 .

that for sufficiently many perturbations, there exists an index $m = m^*$ such that the perturbed quantities, $\widehat{\Sigma}^{[m^*]}$, $\widehat{\gamma}^{[m^*]}$, $\widehat{\mu}_Y^{[m^*]}$, and $\widehat{\mu}^{[m^*]}$, closely retrieve the true population quantities Σ , γ , μ_Y , and μ , respectively; see the Supplementary Material for the theoretical result. Consequently, the optimization problem in (22) becomes nearly equivalent to the population-level problem in (21) when $m = m^*$.

Finally, we construct the perturbed estimator of the treatment effect,

$$\widehat{\tau}^{[m]} = \widehat{\mu}_Y - (\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]}, \quad (27)$$

where $\widehat{\beta}^{[m]}$ is defined in (22). In the above expression, we only replace $\widehat{\mu}^\top \widehat{\beta}$ with the perturbed version $(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]}$ but retain $\widehat{\mu}_Y$ from $\widehat{\tau}$ in (18) since the uncertainty of $\widehat{\mu}_Y$ can be quantified using the asymptotic normality. We mainly use the current perturbation technique to quantify the term $\widehat{\mu}^\top \widehat{\beta}$, which can have a non-regular limiting distribution.

The following decomposition of the estimation error $\widehat{\tau}^{[m]} - \tau^*$ highlights the effectiveness of our proposed perturbation method: for any $m = 1, \dots, M$,

$$\widehat{\tau}^{[m]} - \tau^* = (\widehat{\mu}_Y - \mu_Y) - \left[(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]} - \mu^\top \beta^* \right]. \quad (28)$$

The decomposition (28) reveals that the estimation error $\widehat{\tau}^{[m]} - \tau^*$ consists of two components: the well-behaved $\widehat{\mu}_Y - \mu_Y$ and the perturbation error $(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]} - \mu^\top \beta^*$, which is negligible for certain perturbation $m = m^*$. With high probability, we ensure the existence of such an index m^* which makes $(\widehat{\mu}^{[m^*]})^\top \widehat{\beta}^{[m^*]} \approx \mu^\top \beta^*$ for sufficiently large M ; see the Supplementary Material for details. For such an m^* , we simply need to quantify the asymptotically normal component $\widehat{\mu}_Y - \mu_Y$. We provide numerical evidence for such an m^* existing in Figure 7, where there exists m^* such that $(\widehat{\mu}^{[m^*]})^\top \widehat{\beta}^{[m^*]} - \mu^\top \beta^*$ is much more smaller than $\widehat{\mu}^\top \widehat{\beta} - \mu^\top \beta^*$. This observation supports the intuition behind the decomposition of $\widehat{\tau}^{[m]} - \tau^*$ in (28).

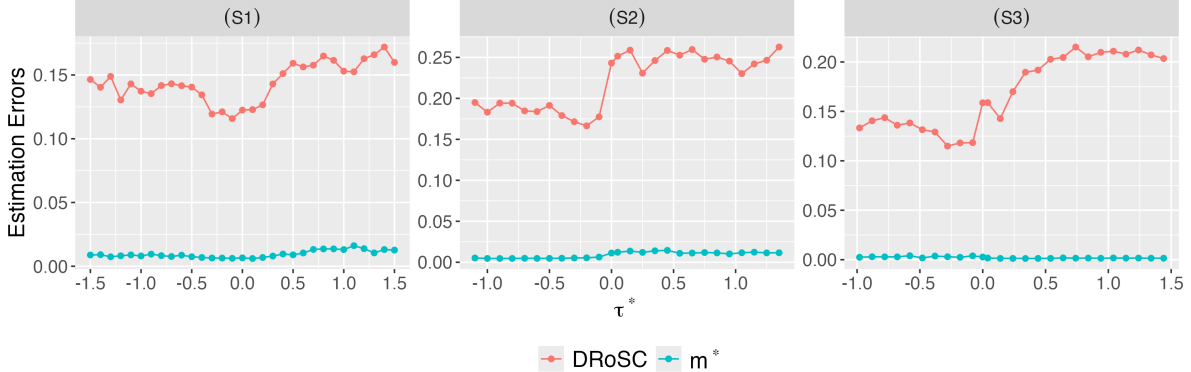


Figure 7: Empirical average of estimation errors under (S1), (S2), and (S3) from Section 7.1, with respect to a range of τ^* values specified in the x -axis. DRoSC and m^* correspond to $|\widehat{\mu}^\top \widehat{\beta} - \mu^\top \beta^*|$ and $\min_{m \in \mathbb{M}} |(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]} - \mu^\top \beta^*|$ where $\mathbb{M}$ is later defined in (29), respectively.

Step 2: Filtering and Aggregation. In the following, we discuss filtering out some inaccurate perturbations. Even though it is impossible to identify the best index m^* , our goal is to retain the perturbation $m = m^*$ and exclude those perturbations that are unlikely to be m^* .

We screen out a small proportion of perturbations if they appear on the tails of the distributions in (24) and (25). To facilitate the discussion, we define normalized resampled statistics as $\hat{T}^{[m]} = \hat{\mathbf{V}}^{-1/2} \hat{U}^{[m]}$ for $m = 1, \dots, M$ where

$$\begin{aligned}\hat{U}^{[m]} &= \left(\hat{\mu}_Y^{[m]} - \hat{\mu}_Y, (\hat{\mu}^{[m]} - \hat{\mu})^\top, (\text{vecl}(\hat{\Sigma}^{[m]} - \hat{\Sigma}))^\top, (\hat{\gamma}^{[m]} - \hat{\gamma})^\top \right)^\top, \\ \hat{\mathbf{V}} &= \text{diag} \left(\hat{\mathbf{V}}_Y, \hat{\mathbf{V}}_\mu + \|\hat{\mathbf{V}}_\mu\|_{\max} \mathbf{I}, \hat{\mathbf{V}}_\Sigma + \|\hat{\mathbf{V}}_\Sigma\|_{\max} \mathbf{I}, \hat{\mathbf{V}}_\gamma + \|\hat{\mathbf{V}}_\gamma\|_{\max} \mathbf{I} \right).\end{aligned}$$

Here, the stacked long vector $\hat{U}^{[m]}$ is the vector of centered m -th perturbed quantities generated from the distributions in (24) and (25). $\hat{\mathbf{V}}$ is a block diagonal matrix containing the covariance matrices of these perturbation-generating distributions, with off-block correlations ignored. Thus, $\hat{T}^{[m]}$ is the resulting vector of normalized deviations between the perturbed quantities and their corresponding estimators.

With $\hat{T}^{[m]}$, we introduce the following index set $\mathbb{M}$ as

$$\mathbb{M} = \left\{ 1 \leq m \leq M : \lambda_{\min}(\hat{\Sigma}^{[m]}) \geq 0, \quad \left\| \hat{T}^{[m]} \right\|_\infty \leq 1.1 z_{\alpha_0/(2p)} \right\}, \quad (29)$$

where z_q is the upper q quantile of the standard normal distribution, $\alpha_0 \in (0, 0.01]$ is a prespecified constant to exclude extreme perturbations on the tail of the distributions (24) and (25), and we introduce the factor 1.1 to adjust for the estimation error. We note that any value greater than 1 can be used. The index set $\mathbb{M}$ in (29) excludes the m -th perturbation if the minimum eigenvalue of $\hat{\Sigma}^{[m]}$ is negative or the maximum of the test statistics exceeds a specified threshold, which is chosen to adjust for multiple comparisons using the Bonferroni correction. Since $\lambda_{\min}(\Sigma) \geq 0$, it is reasonable to filter out $\hat{\Sigma}^{[m]}$ when it starts reporting negative eigenvalues.

Given the significance level $\alpha > \alpha_0$ and for each $m \in \mathbb{M}$, we construct the m -th interval as

$$\text{Int}^{[m]} = \left[\hat{\tau}^{[m]} - z_{\alpha'/2} \hat{\mathbf{V}}_Y^{1/2}, \hat{\tau}^{[m]} + z_{\alpha'/2} \hat{\mathbf{V}}_Y^{1/2} \right], \quad (30)$$

where $\alpha' = \alpha - \alpha_0$, and $\hat{\mathbf{V}}_Y$ is defined in (23) and denotes the estimator of the variance of $\hat{\mu}_Y$. If no feasible solution exists for (22), then $\hat{\tau}^{[m]}$ cannot be obtained. In such cases, we set $\text{Int}^{[m]} = \emptyset$. Finally, we construct the CI for τ^* by the following union aggregation,

$$\text{CI}_\alpha = \bigcup_{m \in \mathbb{M}} \text{Int}^{[m]}, \quad (31)$$

with $\text{Int}^{[m]}$ defined in (30). We refer to CI_α as a confidence interval even though $\cup_{m \in \mathbb{M}} \text{Int}^{[m]}$ may not be an interval. For each $m \in \mathbb{M}$, the interval $\text{Int}^{[m]}$ quantifies the uncertainty of $\hat{\mu}_Y$ at the confidence level α using a standard inference method, while treating $(\hat{\mu}^{[m]})^\top \hat{\beta}^{[m]}$ as being nearly the same as $\mu^\top \beta^*$. By the decomposition (28) and discussion after that, there exists an index $m^* \in \mathbb{M}$ such that $\text{Int}^{[m^*]}$ nearly serves as a level- α' confidence interval for τ^* . However, since the specific identity of such m^* remains unknown, we take the union in (31) to address this uncertainty. We summarize our proposal in Algorithm 2.

Tuning Parameter Selection. We provide details on choosing the tuning parameter ρ_M in a data-dependent way. Since we replace γ and Σ in (14) with their perturbed counterparts $\hat{\gamma}^{[m]}$ and $\hat{\Sigma}^{[m]}$ to construct perturbed $\hat{\Omega}^{[m]}$ in (26), it is necessary to adjust for the error introduced by this substitution. The theorem in the Supplementary Material suggests selecting

$$\rho_M = \frac{C_1}{\sqrt{T_0}} \left[\frac{\log(\min\{T_0, T_1\})}{M} \right]^{1/p}, \quad (32)$$

Algorithm 2 DRoSC inference with perturbation methods

Input: Pre-treatment data $\{Y_{1,t}, X_t\}_{t=1}^{T_0}$; Post-treatment data $\{Y_{1,t}, X_t\}_{t=T_0+1}^T$; Weight shift parameter $\lambda \geq 0$; Sampling size $M = 500$; Significance level $\alpha \in (0, 1)$; Pre-specified constant $\alpha_0 \in (0, 0.01]$; Tuning parameter ρ_M .

Output: Confidence interval CI_α .

- 1: Construct $\widehat{\Sigma}$ and $\widehat{\gamma}$ as in (15), and $\widehat{\mu}_Y$ and $\widehat{\mu}$ as in (17);
 - 2: **for** $m \leftarrow 1, \dots, M$ **do**
 - 3: Sample $\widehat{\Sigma}^{[m]}$ and $\widehat{\gamma}^{[m]}$ as in (24) with $\widehat{\mathbf{V}}_\Sigma, \widehat{\mathbf{V}}_\gamma$ as in (23) and (23);
 - 4: Sample $\widehat{\mu}_Y^{[m]}$ and $\widehat{\mu}^{[m]}$ as in (25) with $\widehat{\mathbf{V}}_Y$ and $\widehat{\mathbf{V}}_\mu$ as in (23);
 - 5: Construct $\widehat{\Omega}^{[m]}$ as in (26);
 - 6: Construct $\widehat{\beta}^{[m]}$ as in (22) and $\widehat{\tau}^{[m]}$ as in (27);
 - 7: Construct $\text{Int}^{[m]}$ as in (30);
 - 8: **end for** ▷ Perturbation
 - 9: Construct the index set $\mathbb{M}$ as in (29); ▷ Filtering
 - 10: Construct CI_α as in (31); ▷ Aggregation
-

for some positive constant $C_1 > 0$. In practice, however, the appropriate value of C_1 is unknown. Drawing an analogy to the selection of the constant C in ρ in (19), we propose to initialize C_1 with a small default value (e.g., $C_1 = 0.01$). However, a small value of ρ_M may lead to feasibility issues due to a similar reason as choosing the tuning parameter ρ in (19). To ensure feasibility of the perturbed optimization problem (22), we require that $\widehat{\Omega}^{[m]}$ in (26) is nonempty. If this condition fails, we regard the solution $\widehat{\beta}^{[m]}$ of (22) as infeasible. We iteratively increase C_1 in (32) by a multiplicative factor of 1.25 until a prespecified proportion of perturbed optimization problems (e.g., 10% by default) are feasible. Numerical studies in Section 7.4 demonstrate robustness to this prespecified proportion. Specifically, compared to our default choice of 10%, setting the proportion to 20% or 30% yields similar performance in terms of CI coverage and length.

Remark 1. Our method is inspired by the repro-sampling method of Xie and Wang (2024) and the resampling idea in Guo (2023b), but it is fundamentally different from both. While Xie and Wang (2024) focused on problems with discrete structures (e.g., the mixture model) and inference after identifying the discrete structures, we address non-regular inference for a continuous parameter, where the non-regular distribution arises from boundary effects and system instability. Guo (2023b) devised a perturbed optimization approach when the population optimization problem is strictly convex, guaranteeing the uniqueness of the perturbed optimizers. In contrast, our method uses a perturbation-based approach for a non-strictly convex optimization problem, where the optimizer $\widehat{\beta}^{[m]}$ from the m -th perturbed problem in (22) may not be unique. As a result, it is possible that $\widehat{\beta}^{[m^*]} \neq \beta^*$ even when the perturbed optimization problem with $m = m^*$ nearly recovers the population optimization problem (21). Nevertheless, the quantity $(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]}$ is uniquely defined for all m , which ensures that $(\widehat{\mu}^{[m^*]})^\top \widehat{\beta}^{[m^*]} - \mu^\top \beta^*$ is negligible for some $m = m^*$. Furthermore, while Guo (2023b) considered the simplex constraint, thus not quantifying the uncertainty of estimating the constraint set, our setting requires quantifying the uncertainty of estimating the constraint set Ω in (14) in a data-dependent way. Incorporating estimation error of the constraint set complicates the theoretical analysis, particularly in establishing the estimator’s convergence rate and validating the proposed statistical inference. We address this challenge and establish a rigorous justification in the theorem in the Supplementary Material.

Remark 2. We remark that the specification of the covariance estimator $\widehat{\mathbf{V}}_\Sigma$, $\widehat{\mathbf{V}}_\gamma$, $\widehat{\mathbf{V}}_Y$, and $\widehat{\mathbf{V}}_\mu$ used in the perturbation procedure (24) and (25) depends on the underlying data-generating mechanism. For example, under the assumption that the pre-treatment and post-treatment data are i.i.d. with fixed N , we use $\widehat{\mathbf{V}}_\Sigma$, $\widehat{\mathbf{V}}_\gamma$, $\widehat{\mathbf{V}}_Y$, and $\widehat{\mathbf{V}}_\mu$ as in (23). For weakly dependent and stationary processes, however, we instead use the Heteroskedasticity and Autocorrelation Consistent (HAC) covariance estimators (Newey and West, 1987; Andrews, 1991) to achieve consistency. In more general settings, however, obtaining consistent covariance estimators can be difficult without the knowledge of the dependence structure of the data. While Assumption 3 shows that consistency of the covariance estimators is sufficient to justify our method, our simulation studies in the Supplementary Material show that even inconsistent covariance estimators can validate the perturbation procedure, provided that a larger number of perturbations M is used and the covariance estimator used in the perturbation-generating distribution is sufficiently large such that near recovery of the true quantities can be generated.

6 Theoretical Justification

In this section, we provide theoretical justification for our proposed framework by (i) establishing the consistency of our estimator $\widehat{\tau}$ introduced in Section 4, and (ii) analyzing the coverage properties of CI_α constructed using our perturbation-based inference method proposed in Section 5.2. To facilitate theoretical analysis, we let T_0 and T_1 grow, and while we can consider growing N , but we focus on a fixed- N regime throughout the paper.

In what follows, we introduce assumptions on the data and error terms in both the pre-treatment and post-treatment periods, which characterize the convergence rates of $\widehat{\tau}$. We begin with the conditions for the pre-treatment period.

Assumption 1. *For the pre-treatment control units' outcomes and error terms $\{X_t, u_t^{(0)}\}_{t=1}^{T_0}$ in (4), there exist positive constants $C_0 > 0$ and $b > 0$ which do not depend on T_0 and N such that*

$$\mathbb{P} \left(\sup_{\beta \in \Delta^N} \left\| \frac{1}{T_0} \sum_{t=1}^{T_0} \left[X_t u_t^{(0)} + (X_t X_t^\top - \mathbb{E} X_t X_t^\top)(\beta^{(0)} - \beta) \right] \right\|_\infty \leq \frac{C_0 [\log(\max\{T_0, N\})]^{\frac{1+b}{2b}}}{\sqrt{T_0}} \right) \rightarrow 1$$

as $T_0 \rightarrow \infty$.

Assumption 1 requires that the empirical averages of $X_t u_t^{(0)}$ and the deviation of the sample covariance matrix from its expectation, remain uniformly well controlled over $\beta \in \Delta^N$. Since $\mathbb{E} X_t u_t^{(0)} = 0$, Assumption 1 ensures that these empirical fluctuations vanish to zero as T_0 grows, so that the $\widehat{\Omega}$ behaves similarly to its population counterpart, Ω . If $\{X_t, u_t^{(0)}\}_{t=1}^{T_0}$ are i.i.d. with fixed N , this assumption holds for all $b > 0$. More generally, it holds when the pre-treatment data are β -mixing with exponential decay, where b is related to the order of β -mixing coefficients (see, Chernozhukov et al., 2021, Lemma H.8). A similar condition is used in the literature to control prediction error of the SC estimator in the absence of weight shifts (e.g., Chernozhukov et al., 2021; Ben-Michael et al., 2021).

Next, we introduce an assumption for the post-treatment error terms. For the detailed assumption, we define error terms of control units' outcomes for post-treatment periods $\nu_t = X_t - \mathbb{E} X_t$ for $t = T_0 + 1, \dots, T$.

Assumption 2. The error terms $\{\epsilon_t\}_{t=T_0+1}^T$ satisfy that $T_1^{-1/2} \sum_{t=T_0+1}^T \epsilon_t = O_p(1)$ as $T_1 \rightarrow \infty$, where $\epsilon_t = (\nu_t^\top, v_t, u_t^{(1)})^\top$, and v_t and $u_t^{(1)}$ are defined in (2) and (4), respectively. Furthermore, $\mu = T_1^{-1} \sum_{t=T_0+1}^T \mathbb{E}X_t$ is bounded as $T_1 \rightarrow \infty$.

Assumption 2 holds for i.i.d. post-treatment data. Assumption 2 may hold under more general settings with dependent structures, such as strong mixing, provided that suitable conditions are satisfied (see, e.g., Billingsley, 2017, Theorem 27.4). A similar assumption regarding the post-treatment data is used in the SC literature (e.g., Li, 2020).

The following theorem establishes the convergence rate for the DRoSC estimator $\hat{\tau}$ defined in (18), where we divide the presentation into two cases based on $\lambda > 0$ and $\lambda = 0$.

Theorem 3. Suppose Assumptions 1 and 2 hold, and that the tuning parameter ρ used in (15) satisfies $\rho = C[\log(\max\{T_0, N\})]^{1+b/2b}/\sqrt{T_0}$ with some positive constant C which satisfies $C \geq C_0$ from Assumption 1. Then, for $\hat{\tau}$ defined in (18), the following holds with $b > 0$ from Assumption 1:

(i) Case of $\lambda > 0$:

$$\lim_{T_0, T_1 \rightarrow \infty} \mathbb{P} \left(|\hat{\tau} - \tau^*| \lesssim \left[\frac{[\log(\max\{T_0, N\})]^{1+b/2b}}{\sqrt{T_0} \cdot \lambda} \right]^{1/2} + \left[\frac{1}{\sqrt{T_1}} \right]^{1/2} \right) = 1. \quad (33)$$

(ii) Case of $\lambda = 0$: under the additional condition of $\lambda_{\min}(\Sigma) > 0$,

$$\lim_{T_0, T_1 \rightarrow \infty} \mathbb{P} \left(|\hat{\tau} - \tau^*| \lesssim \left[\frac{[\log(\max\{T_0, N\})]^{1+b/2b}}{\sqrt{T_0} \cdot \lambda_{\min}(\Sigma)/\sqrt{N}} \right]^{1/2} + \left[\frac{1}{\sqrt{T_1}} \right]^{1/2} \right) = 1. \quad (34)$$

For the $\lambda > 0$ case, the convergence rate (33) of $|\hat{\tau} - \tau^*|$ in Theorem 3 comprises two components: the first term, $([\log(\max\{T_0, N\})]^{1+b/2b}/(\sqrt{T_0} \cdot \lambda))^{1/2}$, reflects the error in estimating the uncertainty class Ω by $\hat{\Omega}$; the second term, $(1/\sqrt{T_1})^{1/2}$, corresponds to the error in estimating the objective function $\tau(\beta)$ by $\hat{\mu}_Y - \hat{\mu}^\top \beta$. The rate in (33) also covers limiting case where $\lambda \rightarrow 0$. However, to guarantee the consistency of DRoSC estimator $\hat{\tau}$ as $T_0, T_1 \rightarrow \infty$, we still require a sufficiently large λ such that $[\log(\max\{T_0, N\})]^{1+b/2b}/(\sqrt{T_0} \cdot \lambda) \rightarrow 0$. We shall remark that the convergence rate of $|\hat{\tau} - \tau^*|$ is slower than $1/\sqrt{\min\{T_0, T_1\}}$ mainly due to the reason that the population optimization problem (12) is only convex instead of strictly convex. Specifically, the matrix $\mu\mu^\top$ in (12) is only of rank one, leading to (12) being in the form of a degenerate quadratic optimization. This degeneracy precludes the use of standard M -estimation theory and, as a result, prevents achieving the usual parametric rate $1/\sqrt{\min\{T_0, T_1\}}$ for $|\hat{\tau} - \tau^*|$. This phenomenon is analogous to the slow convergence rates in high-dimensional regression problems: when there is a lack of restricted eigenvalue or restricted strong convexity conditions, it is challenging to establish the fast parametric rate but the literature can establish the slow convergence rate for the prediction problem (see, e.g., Bühlmann and van de Geer, 2011; Wainwright, 2019).

For the case of $\lambda = 0$, the additional condition $\lambda_{\min}(\Sigma) > 0$ in Theorem 3(ii) is required to ensure that Ω contains a unique weight; this uniqueness is necessary for the convergence of $\hat{\Omega}$ to Ω when $\lambda = 0$. Thus, the theorem precludes non-unique SC weights when no weight shift is assumed by $\lambda = 0$. Theorem 3(ii) also covers limiting case where $\lambda_{\min}(\Sigma) \rightarrow 0$ but it requires $\lambda_{\min}(\Sigma)$ to satisfy $[\log(\max\{T_0, N\})]^{1+b/2b}/(\sqrt{T_0} \cdot \lambda_{\min}(\Sigma)/\sqrt{N}) \rightarrow 0$, as similar to (33).

Finally, we provide theoretical justification for the validity of our perturbation-based inference method. We impose the following assumption on the limiting distribution of $\text{vecl}(\widehat{\Sigma})$, $\widehat{\gamma}$, $\widehat{\mu}_Y$, and $\widehat{\mu}$, as well as on the consistency of the corresponding covariance estimators.

Assumption 3. $\text{vecl}(\widehat{\Sigma})$, $\widehat{\gamma}$, $\widehat{\mu}_Y$, and $\widehat{\mu}$ admit the following asymptotic distributions:

$$\begin{aligned} \mathbf{V}_\Sigma^{-1/2} \left(\text{vecl}(\widehat{\Sigma}) - \text{vecl}(\Sigma) \right) &\xrightarrow{d} \mathcal{N}(0, \mathbf{I}), \quad \mathbf{V}_\gamma^{-1/2} (\widehat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}), \quad \text{as } T_0 \rightarrow \infty, \\ \text{and } \mathbf{V}_Y^{-1/2} (\widehat{\mu}_Y - \mu_Y) &\xrightarrow{d} \mathcal{N}(0, 1), \quad \mathbf{V}_\mu^{-1/2} (\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}) \quad \text{as } T_1 \rightarrow \infty, \end{aligned} \quad (35)$$

for some positive constant V_Y and positive definite matrices $\mathbf{V}_\Sigma$, $\mathbf{V}_\gamma$, and $\mathbf{V}_\mu$, where there exists a constant $C > 0$ such that

$$\lim_{T_0 \rightarrow \infty} \mathbb{P}(\max\{\|T_0 \mathbf{V}_\Sigma\|_2, \|T_0 \mathbf{V}_\gamma\|_2\} \leq C) = \lim_{T_1 \rightarrow \infty} \mathbb{P}(\max\{T_1 V_Y, \|T_1 \mathbf{V}_\mu\|_2\} \leq C) = 1.$$

Furthermore, the rescaled covariance estimators $\widehat{\mathbf{V}}_\Sigma$, $\widehat{\mathbf{V}}_\gamma$, $\widehat{\mathbf{V}}_Y$, and $\widehat{\mathbf{V}}_\mu$ defined in (23) satisfy consistency:

$$\begin{aligned} \|T_0(\widehat{\mathbf{V}}_\Sigma - \mathbf{V}_\Sigma)\|_2 &\xrightarrow{p} 0, \quad \|T_0(\widehat{\mathbf{V}}_\gamma - \mathbf{V}_\gamma)\|_2 \xrightarrow{p} 0, \quad \text{as } T_0 \rightarrow \infty, \\ |T_1(\widehat{\mathbf{V}}_Y - V_Y)| &\xrightarrow{p} 0, \quad \|T_1(\widehat{\mathbf{V}}_\mu - \mathbf{V}_\mu)\|_2 \xrightarrow{p} 0, \quad \text{as } T_1 \rightarrow \infty. \end{aligned}$$

Assumption 3 implies that the asymptotic normality of simple estimators for population quantities $\text{vecl}(\widehat{\Sigma})$, $\widehat{\gamma}$, $\widehat{\mu}_Y$, and $\widehat{\mu}$, as well as the consistency of the covariance estimators $\widehat{\mathbf{V}}_\Sigma$, $\widehat{\mathbf{V}}_\gamma$, $\widehat{\mathbf{V}}_Y$, and $\widehat{\mathbf{V}}_\mu$. The central limit theorem (35) in Assumption 3 is satisfied when the pre-treatment and post-treatment data are i.i.d. in the fixed N regime, and more generally, under α -mixing and stationarity (e.g., Billingsley, 2017, Theorem 27.4). For consistent estimation of the covariance, the choice of estimator should reflect the data regime: the covariance estimators in (23) are appropriate under i.i.d. sampling, whereas HAC estimators can be adopted to accommodate weak dependence and stationarity (Newey and West, 1987; Andrews, 1991).

We introduce the following function $\text{err}(\cdot)$ depending on the resampling size M to quantify the effect of the perturbation step,

$$\text{err}(M) = \frac{1}{2} \left[\frac{2 \log(\min\{T_0, T_1\})}{c^*(\alpha_0) \cdot M} \right]^{1/p}, \quad (36)$$

where $\alpha_0 \in (0, 0.01]$ is a pre-specified constant used to construct $\mathbb{M}$ in (29), and $c^*(\alpha_0)$ is defined in the Supplementary Material. Since $c^*(\alpha_0)$ is a constant that only depends on the pre-specified α_0 , $\text{err}(M)$ is of order $[\log(\min\{T_0, T_1\})/M]^{1/p}$, which converges to zero as M goes to infinity. For the case of $\lambda > 0$, we establish in the Supplementary Material that, with high probability,

$$\min_{m \in \mathbb{M}} \left| (\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]} - \mu^\top \beta^* \right| \lesssim \left[\frac{\text{err}(M)}{\sqrt{\min\{T_0, T_1\}} \cdot \lambda} \right]^{1/2}, \quad (37)$$

where a similar result can be obtained for the case $\lambda = 0$; see the Supplementary Material for details. We define m^* to be one index attaining $\min_{m \in \mathbb{M}} |(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]} - \mu^\top \beta^*|$. Importantly, the above upper bound states that, with a sufficiently large M , the term $|(\widehat{\mu}^{[m]})^\top \widehat{\beta}^{[m]} - \mu^\top \beta^*|$ goes to zero even for a fixed T_0 and T_1 . This means that there exists one perturbation such that the uncertainty from the non-regular component disappears.

The following theorem presents the main result of our perturbation-based inference procedure, establishing the coverage property of the proposed CI.

Theorem 4. Suppose Assumptions 2 and 3 hold, and the tuning parameter ρ_M used in (26) satisfies $\rho_M = C_1[\log(\min\{T_0, T_1\})/M]^{1/p}/\sqrt{T_0}$ with some positive constant C_1 which satisfies $C_1 \geq (2/c^*(\alpha_0))^{1/p}$, where $c^*(\alpha_0)$ is used in (36) for $\alpha_0 \in (0, 0.01]$. For $\alpha \in (\alpha_0, 1)$, CI_α defined in (31) satisfies the following coverage property:

(i) Case of $\lambda > 0$:

$$\liminf_{T_0, T_1 \rightarrow \infty} \liminf_{M \rightarrow \infty} \mathbb{P}(\tau^* \in \text{CI}_\alpha) \geq 1 - \alpha. \quad (38)$$

(ii) Case of $\lambda = 0$: under the additional assumption of $\lambda_{\min}(\Sigma) > 0$, (38) still holds.

We note that (38) establishes only a one-sided coverage guarantee, as our proposed perturbation-based method involves taking a union over $\mathbb{M}$. Recalling the decomposition (28), the estimation error $\hat{\tau}^{[m]} - \tau^*$ depends on the regular uncertainty from $\hat{\mu}_Y - \mu_Y$, which is asymptotically normal with a $1/\sqrt{T_1}$ convergence rate, and the non-regular term $(\hat{\mu}^{[m]})^\top \hat{\beta}^{[m]} - \mu^\top \beta^*$. Since $\text{err}(M) \rightarrow 0$ as $M \rightarrow \infty$, the error bounds (37) of the non-regular term can be made negligible relative to the $1/\sqrt{T_1}$ uncertainty for some $m = m^*$ by choosing a sufficiently large M . Thus, the non-regular term does not affect the limiting distribution of $\hat{\mu}_Y - \mu_Y$, thereby guaranteeing the validity of the coverage property (38).

Now we turn to the length of our proposed CI. Prior to presenting the main theoretical result, we define a refined index set $\tilde{\mathbb{M}}$ based on $\mathbb{M}$ in (29), which will be used in the following theoretical analysis of CI length. Specifically, we define $\tilde{\mathbb{M}}$ as

$$\tilde{\mathbb{M}} = \mathbb{M} \cap \left\{ 1 \leq m \leq M : \hat{\Omega}^{[m]}(0) \text{ is non-empty} \right\}. \quad (39)$$

where $\hat{\Omega}^{[m]}(\lambda)$ is defined in (26). The rationale for this refinement is as follows. The best perturbation index m^* should yield a perturbed uncertainty class $\hat{\Omega}^{[m^*]}(0)$ that nearly recovers the corresponding true uncertainty class $\Omega(0)$. Since $\Omega(0)$ contains the true pre-treatment weight $\beta^{(0)}$, $\hat{\Omega}^{[m^*]}(0)$ should also contain $\beta^{(0)}$, ensuring that $\hat{\Omega}^{[m^*]}(0)$ is non-empty. Thus, we filter out indices that do not satisfy the condition in (39), as they are unlikely to correspond to the best index m^* . We note that $m^* \in \tilde{\mathbb{M}}$, so the CI constructed using $\tilde{\mathbb{M}}$ retains the same coverage property established in Theorem 4.

The following theorem presents our main result on the length of the proposed CI.

Theorem 5. Suppose Assumptions 2 and 3 hold, and the tuning parameter ρ_M used in (26) satisfies $\rho_M = C_1[\log(\min\{T_0, T_1\})/M]^{1/p}/\sqrt{T_0}$ with some positive constant C_1 . For $\alpha \in (\alpha_0, 1)$, Suppose also that CI_α in (31) is constructed using the refined index set $\tilde{\mathbb{M}}$ in (39). Then the length of CI_α , denoted by $\mathbf{L}(\text{CI}_\alpha)$, satisfies the following precision property:

(i) Case of $\lambda > 0$:

$$\liminf_{T_0, T_1 \rightarrow \infty} \liminf_{M \rightarrow \infty} \mathbb{P} \left(\mathbf{L}(\text{CI}_\alpha) \lesssim \left[\frac{1}{\sqrt{T_0} \cdot \lambda} \right]^{1/2} + \left[\frac{1}{\sqrt{T_1}} \right]^{1/2} \right) \geq 1 - \alpha_0. \quad (40)$$

(ii) Case of $\lambda = 0$: under the additional assumption of $\lambda_{\min}(\Sigma) > 0$,

$$\liminf_{T_0, T_1 \rightarrow \infty} \liminf_{M \rightarrow \infty} \mathbb{P} \left(\mathbf{L}(\text{CI}_\alpha) \lesssim \left[\frac{\sqrt{N}}{\sqrt{T_0} \cdot \lambda_{\min}(\Sigma)} \right]^{1/2} + \left[\frac{1}{\sqrt{T_1}} \right]^{1/2} \right) \geq 1 - \alpha_0. \quad (41)$$

Consistent with Theorem 3, the length of our proposed CI, $\mathbf{L}(\text{CI}_\alpha)$, is characterized separately for the cases $\lambda > 0$ and $\lambda = 0$ in Theorem 4, and it also fails to attain the rate of $1/\sqrt{\min\{T_0, T_1\}}$ due to the lack of strict convexity in the optimization problem (12). It is unclear whether the constructed confidence interval achieves the optimal precision property. However, we remark that the validity of CI_α in terms of coverage still holds and it does not rely on the asymptotic normality of $\hat{\tau}$, which may fail due to non-regularity and instability. In Figure 9 in Section 7.3, we investigate the finite-sample performance of our proposed CI and compare it to methods based on asymptotic normality. We find that the length of our proposed CI is slightly larger, but overall comparable to the oracle benchmark.

7 Simulation Studies

In this section, we evaluate the performance of the DRoSC estimation and inference procedures through numerical studies. Section 7.1 describes the simulation settings, Section 7.2 compares the DRoSC and SC estimators, and Section 7.3 assesses the proposed confidence intervals. Finally, Section 7.4 examines the robustness of our inference method with respect to the choice of tuning parameters.

7.1 Simulation Setup

In the following, we detail the simulation design used to assess the accuracy of our proposed estimator in Section 7.2 and the coverage properties of our proposed inference procedure in Section 7.3. Throughout the simulations, we generate the control units' outcomes according to a stationary AR(1) model with AR(1) coefficient $\phi \in [0, 1)$: the pre-treatment control units' outcome observations $\{X_t\}_{t=1}^{T_0}$ are generated with mean μ_0 and equi-correlation covariance matrix $\Sigma_0 = (1 - \rho_0)\mathbf{I}_N + \rho_0\mathbf{1}_N\mathbf{1}_N^\top$, while the post-treatment control units' outcome observations $\{X_t\}_{t=T_0+1}^T$ are generated with mean μ and identity covariance matrix $\mathbf{I}_N$. We present the mathematical formulation of the data-generating model for $\{X_t\}_{t=1}^T$ used in our simulations in Supplementary Material

The treated unit's potential outcomes of receiving the control are generated according to the model in (4), with $\beta^{(0)} = (1/3, 1/3, 1/3, 0, \dots, 0)^\top$ and $u_t^{(0)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ for $t = 1, \dots, T_0$ and $u_t^{(1)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ for $t = T_0 + 1, \dots, T$. For simplicity, we generate the treated unit's potential outcomes of receiving treatment as $Y_{1,t}^{(1)} - Y_{1,t}^{(0)} = \tau + v_t$, where $v_t \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 0.25^2)$ for $t = T_0 + 1, \dots, T$. This implies $\tau_t = \tau$ for $t = T_0 + 1, \dots, T$, and thus $\bar{\tau} = \tau$. We choose the different value of the time-averaged ATT $\bar{\tau}$ among $\{-1.5, -1.4, \dots, 1.4, 1.5\}$, and consider combinations of $(T_0, T_1) \in \{25, 50\} \times \{25, 50\}$, while fixing the number of control units at $N = 10$. The AR(1) coefficient ϕ is varied over $\{0, 0.5\}$.

We consider three settings by varying μ_0 , μ , ρ_0 , and $\beta^{(1)}$. We generate the setting (S1) with Conditions (E1) and (E2) satisfied.

$$(S1) \quad \mu_0 = \mu = (0.8, 1.2, \dots, 0.8, 1.2)^\top, \rho_0 = 0.25, \text{ and } \beta^{(1)} = \beta^{(0)}.$$

In this setting, the control units are not highly correlated and there exist no weight shifts, so that $\bar{\tau}$ is identifiable by the SC method. In the additional two settings, we introduce violations of (E1) or (E2).

$$(S2) \quad \text{We consider a mean shift with } \mu = \mu_0 + (0.6, 0.4, 0.2, \mathbf{0}_{N-3})^\top, \text{ a small weight shift via } \beta^{(1)} = \beta^{(0)} + 0.05 \cdot (-1, \mathbf{0}_{N-2}^\top, 1)^\top, \text{ and induce high correlations among control units by setting } \rho_0 = 0.95.$$

(S3) We consider a mean shift with $\mu_0 = \mu = \mathbf{1}_N + N^{-1}(1, \dots, N)^\top$, a large weight shift via $\beta^{(1)} = \beta^{(0)} + 0.2 \cdot (-\mathbf{1}_3^\top, \mathbf{0}_{N-6}^\top, \mathbf{1}_3^\top)^\top$, and consider settings with low correlations by setting $\rho_0 = 0.25$.

In the main paper, we present results for the case $\phi = 0$, corresponding to i.i.d. data, with $T_0 = 25$ and $T_1 = 25$. Since we report only a subset of settings, we provide additional simulation results with $T_1 = 50$ and $\phi = 0.5$ in the Supplementary Material.

7.2 Estimation

In this section, we evaluate the numerical performance of our proposed estimator $\hat{\tau}$ for τ^* , defined in (18) under settings (S1), (S2), and (S3), as described in Section 7.1. We implement our proposed DRoSC estimation procedure summarized in Algorithm 1. For (S1), we set $\lambda = 0$. For (S2) and (S3), we choose $\lambda = \|\Sigma(\beta^{(1)} - \beta^{(0)})\|_\infty$, where $\Sigma = \Sigma_0 + \mu_0\mu_0^\top$. We implement 500 simulations for each setting and compare the performance with the SC estimator $\hat{\tau}^{\text{SC}}$ for $\bar{\tau}$, defined in (6). We present the results with $\bar{\tau} = -1.5, -1.2$, and -1 for (S1), (S2), and (S3) using violin plots in Figure 8, with additional results with $T_1 = 50$ and $\phi = 0.5$ in the Supplementary Material.

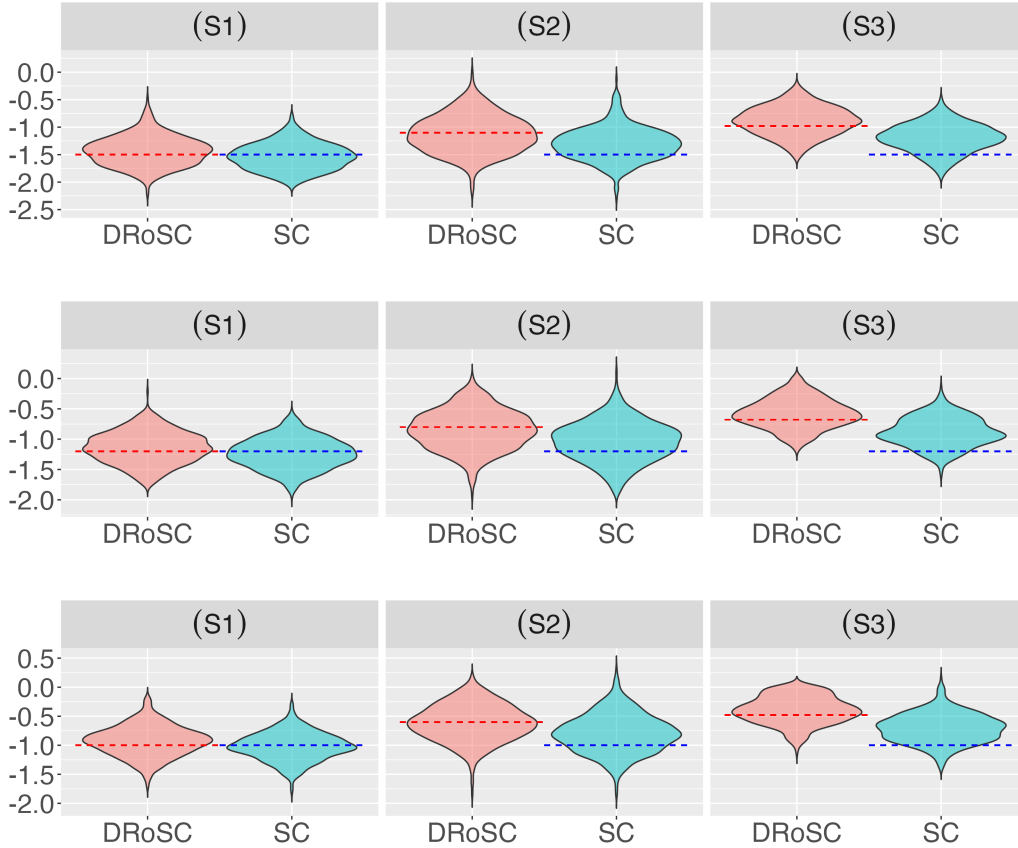


Figure 8: Simulation studies from settings (S1), (S2), and (S3). The figures, from top to bottom, correspond to $\bar{\tau} = -1.5, -1.2$, and -1 respectively. DRoSC and SC in x -axis denote the estimators (18) and (6) from our method and the SC method respectively. The red and blue dashed lines denote τ^* and $\bar{\tau}$ respectively.

In (S1), we have $\bar{\tau} = \tau^*$ because $\lambda = 0$ and $\lambda_{\min}(\Sigma) > 0$, making $\bar{\tau}$ identifiable; hence, both $\hat{\tau}^{\text{SC}}$ and $\hat{\tau}$ share the same target, that is, in Figure 8, the blue dashed line ($\bar{\tau}$) and the red dashed line (τ^*) coincide in (S1). In this setting, τ^* is the same as $\bar{\tau}$ and both $\hat{\tau}^{\text{SC}}$ and $\hat{\tau}$ are consistent. This is illustrated in Figure 8, where the center of each violin plot aligns with its corresponding target, indicated by the dashed lines. In contrast, for settings (S2) and (S3), the identification conditions (E1) and (E2) are violated, making $\bar{\tau}$ unidentifiable and resulting in $\tau^* \neq \bar{\tau}$, as illustrated in Figure 8 by the discrepancy between the red and blue dashed lines. Since $\bar{\tau}$ is not identifiable, $\hat{\tau}^{\text{SC}}$ is no longer consistent for $\bar{\tau}$, as illustrated in the violin plots where the empirical distribution of the estimates deviates from the target value $\bar{\tau}$ (blue dashed lines). This misalignment highlights that, under these settings, the SC approach cannot consistently estimate $\bar{\tau}$ mainly due to the non-identifiability issue of the causal estimand $\bar{\tau}$. However, our proposed estimator consistently estimates τ^* (red dashed lines) for all settings even when the identification conditions (E1) and (E2) are violated.

7.3 Inference

In this section, we evaluate the performance of our perturbation-based inference method introduced in Section 5.2. We set the confidence level to $1 - \alpha = 0.95$, and conduct 500 simulations for each setting described in Section 7.1. Note that for different settings, even though the values of $\bar{\tau}$ are different, their corresponding τ^* can be the same, for example, $\tau^* = 0$ when $\bar{\tau} \in \{-0.4, \dots, 0.1\}$ under (S2). When there are multiple values of τ^* , we report the minimum coverage and the maximum length of the confidence intervals. Our procedure, summarized in Algorithm 2, is implemented with $M = 500$. We denote the CI constructed via our proposed method in (31) as Proposed.

We compare the Proposed CI with CIs based on normality assumptions. For the estimator $\hat{\tau}$ defined in (18), the assumption of the normality-based CIs is that

$$(\hat{\tau} - \tau^*)/\text{SE}(\hat{\tau}) \xrightarrow{d} \mathcal{N}(b^*, 1), \quad (42)$$

where $\text{SE}(\hat{\tau})$ denotes the standard error of $\hat{\tau}$, and b^* represents the associated bias. When $\hat{\tau}$ satisfies (42) with $b^* = 0$, a valid oracle confidence interval based on the asymptotic normality is

$$\left[\hat{\tau} - z_{\alpha/2} \widehat{\text{SE}}(\hat{\tau}), \hat{\tau} + z_{\alpha/2} \widehat{\text{SE}}(\hat{\tau}) \right], \quad (43)$$

where $\widehat{\text{SE}}(\hat{\tau})$ is the empirical standard error of $\hat{\tau}$ across the 500 simulation replicates. We refer to such an oracle confidence interval as the Normality CI. However, when there exists a bias component with $b^* \neq 0$, we follow (6) of Armstrong et al. (2023) and construct an oracle bias-aware (OBA) CI for τ^* as the new benchmark. This interval uses the oracle knowledge of the bias $|\mathbb{E}\hat{\tau} - \tau^*|$ as a rescaled term of the bias b^* ,

$$[\hat{\tau} - \chi_{\alpha}^*, \hat{\tau} + \chi_{\alpha}^*], \quad \text{with} \quad \chi_{\alpha}^* = \widehat{\text{SE}}(\hat{\tau}) \sqrt{\text{cv}_{\alpha} \left(|\mathbb{E}\hat{\tau} - \tau^*|^2 / \widehat{\text{SE}}(\hat{\tau})^2 \right)}, \quad (44)$$

where $\text{cv}_{\alpha}(B^2)$ denotes the $1 - \alpha$ quantile of the non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter B^2 , $\widehat{\mathbb{E}}\hat{\tau}$ is the empirical mean of $\hat{\tau}$ across the 500 simulation replicates, and $\widehat{\text{SE}}(\hat{\tau})$ is the same as used in (43). The OBA CI leverages oracle knowledge of the bias and is not a practical procedure. However, we adopt it as the benchmark as it reflects the best possible interval when there is oracle information of the bias and standard error of $\hat{\tau}$.

We present inference results with empirical coverages and empirical mean lengths of CIs in Figure 9. In (S1), the Normality CI achieves near-nominal coverage overall but under-covers for some values of τ^* (e.g., 0.5 and 1.4). This occurs even though $\bar{\tau}$ is identifiable and coincides with τ^* in (S1), since boundary effects can induce non-regularity in $\hat{\mu}^\top \hat{\beta}$, leading to inference challenges. Notably, in this case our DRoSC method coincides with the standard SC method, so the observed under-coverage reflects an inference challenge inherent to SC itself, as discussed in Cattaneo et al. (2021). In contrast, the OBA and Proposed CIs maintain valid coverage, with the Proposed CI slightly more conservative than the oracle benchmark.

Under (S2) and (S3), the Normality CI exhibits under-coverage. In (S2), coverage is close to 0.95 when τ^* is negative and away from zero, but drops to around 0.9 for positive τ^* and falls below 0.9 near zero, reflecting the non-regularity discussed in Section 5.1. The OBA and Proposed CIs maintain uniformly valid coverage, with the Proposed CI somewhat conservative but still comparable in length to the oracle benchmark OBA. In (S3), a similar pattern holds: the Normality CI misses nominal coverage when $\tau^* = 0$ and $\tau^* > 0.5$, while OBA and Proposed CIs remain valid. Although the Proposed CI is slightly more conservative and modestly longer, its performance often matches that of OBA, which requires oracle knowledge of the bias of the DRoSC estimator.

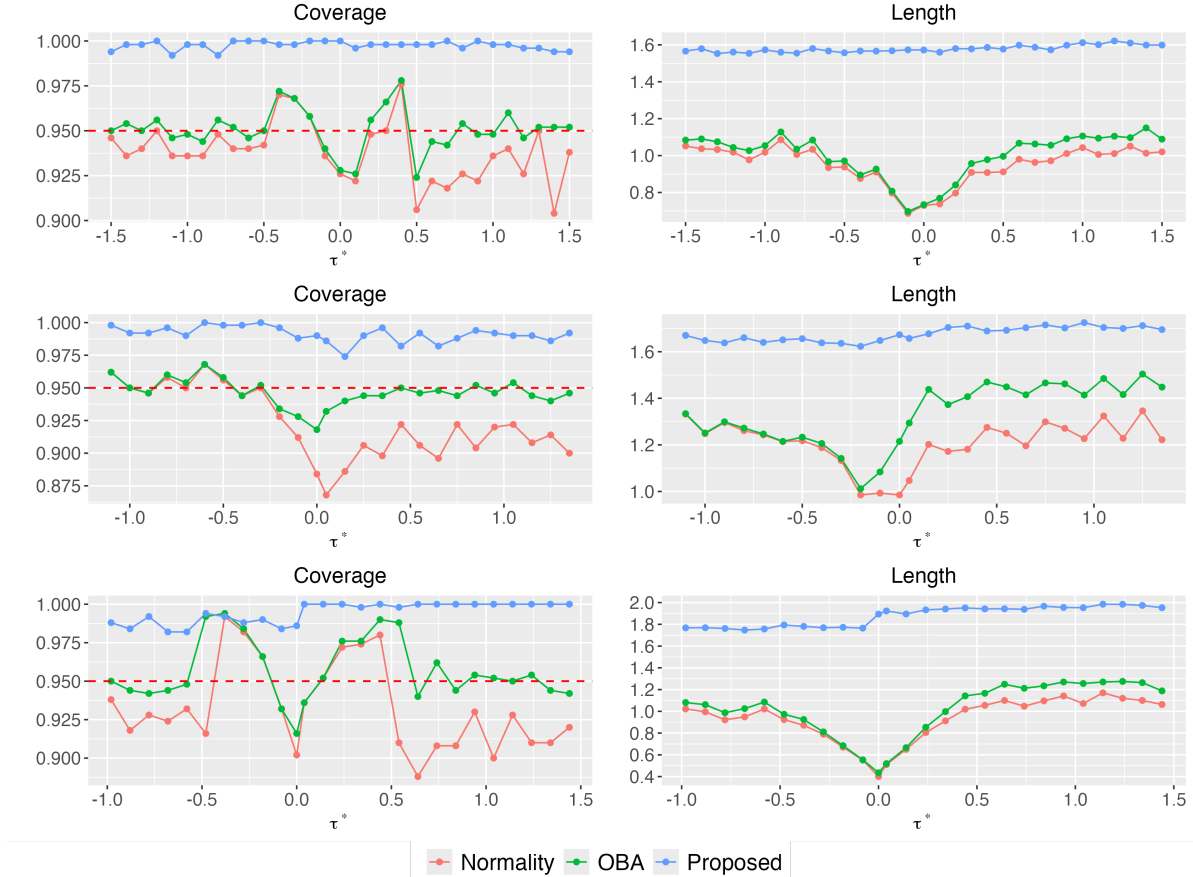


Figure 9: Empirical coverage and interval lengths for settings (S1), (S2), and (S3) described in Section 7.3. The panels from top to bottom respectively correspond to (S1), (S2), and (S3), where x -axis plots the values of τ^* . **Normality** refers to the Normality CI defined in (43), **OBA** denotes the OBA CI in (44), and **Proposed** corresponds to the proposed CI in (31).

7.4 Sensitivity to Tuning Parameters

Our perturbation-based inference method involves a tuning parameter, defined as the proportion of feasible solutions $\widehat{\beta}^{[m]}$ obtained from the perturbed optimization problem (22) across M perturbations. We set the default threshold to 10%. Since this choice may influence the coverage and precision of the confidence interval in (31), we empirically examine the sensitivity of the results to this proportion. The default threshold requires that at least 10% of the M perturbations yield feasible solutions, but we also consider alternative thresholds of 20% and 30% in this section. Consistent with the previous inference simulation in Section 7.3, we set $M = 500$. As shown in Figure 10, the average length of the proposed CI varies slightly with the choice of threshold in settings (S1) and (S2), while the empirical coverage remains relatively stable. In contrast, both coverage and length remain largely unchanged in (S3). This difference arises due to the value of λ used in each simulation. Specifically, we recall that we set $\lambda = \|\Sigma(\beta^{(1)} - \beta^{(0)})\|_\infty$ in (S2) and (S3). Since the weight shifts are small in (S2), the corresponding λ is also small and it is even zero in (S1). In contrast, (S3) involves large weight shifts, resulting in a larger λ . This larger λ results in a greater number of feasible solutions even with a small tuning parameter ρ_M , thereby making the effect of the threshold choice less pronounced in this setting. This pattern is evident in the rightmost panel of Figure 10: the proportion of feasible solutions increases with the threshold in (S1) and (S2), but shows little change in (S3).

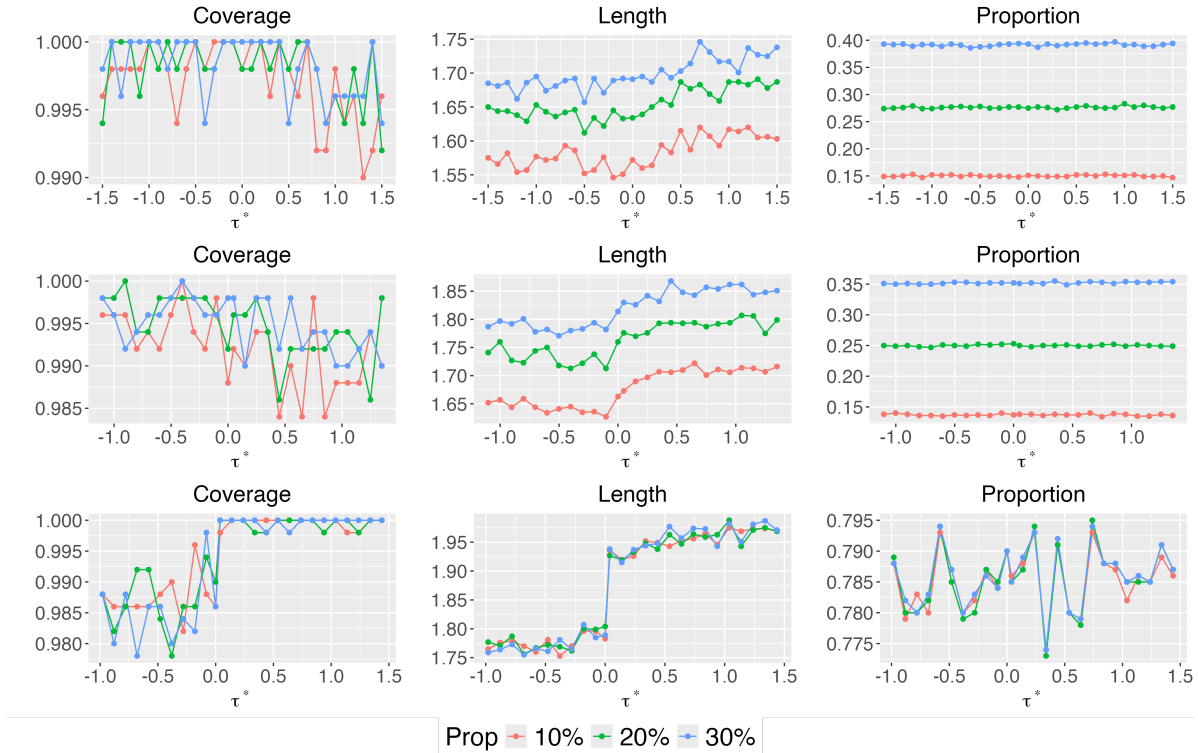


Figure 10: Sensitivity of coverage and average length of the proposed CI to varying proportions of feasible solutions. The leftmost and middle panels denote the empirical coverages and means of lengths of confidence intervals. The rightmost panel displays the empirical proportion of feasible solutions for each threshold. x -axes in all panels denote τ^* . **Prop** denotes the threshold for the proportion of feasible solutions among the $M = 500$ perturbations.

8 Real Data Applications

In this section, we reanalyze the Basque Country case study (Abadie and Gardeazabal, 2003) using SC and our proposed method. The original study examined the economic impact of terrorism in the Basque Country, where the outcome of interest was per capita GDP for the Basque Country and $N = 16$ other Spanish regions (e.g., Madrid, Balears, and Rioja) observed from 1955 to 1997 ($T = 43$). Since terrorism occurred only in the Basque Country, we consider it the treated unit, with the remaining regions serving as controls. The pre-treatment period covers 1955–1969 ($T_0 = 15$), before the first wave of terrorist activity, and the post-treatment period spans 1970–1997 ($T_1 = 28$). To assess the economic impact of terrorism in the Basque Country between 1970 and 1997, Abadie and Gardeazabal (2003) introduced the SC method to construct a counterfactual trajectory of GDP in the absence of terrorism and compare it with the observed GDP. The counterfactual was obtained as a weighted average of control units, with weights selected to match the Basque Country on pre-treatment outcomes and key economic and demographic covariates. For details of the original implementation, see Abadie and Gardeazabal (2003).

As discussed in Section 2.3, however, the control units are highly correlated, and the relationship between the treated unit (Basque Country) and the control units may shift after the onset of terrorism. In particular, the set of weights that best approximate the Basque Country before terrorism may no longer provide a valid representation in the post-treatment period, effectively resulting in a shift of the optimal weights. Both features raise concerns about the stability of the SC estimator and the reliability of its causal conclusions. To address these concerns, we apply our proposed DRoSC method to the Basque data. While the time-averaged ATT is not point-identifiable under weight shifts, we can still estimate and make inference about its conservative proxy, the weight-robust treatment effect. To examine how the result changes by increasing degrees of weight shift, we vary the weight-shift parameter λ over $\{0, 0.001, \dots, 0.06\}$ and, for each value, conduct estimation and inference with confidence level $\alpha = 0.05$ and $M = 500$ following Algorithms 1 and 2, respectively. For comparison, we also report point estimates (6) from the standard SC method based on outcomes only, while the original study in Abadie and Gardeazabal (2003) included additional covariates.

We present the estimation and inference results on the left panel of Figure 11. When $\lambda = 0$, our estimator yields $\hat{\tau} \approx -0.76$, compared to the SC estimate $\hat{\tau}^{\text{SC}} \approx -0.89$, providing a more conservative estimate even without allowing post-treatment weight shifts. As λ increases, $\hat{\tau}$ rises monotonically, reaching zero at $\lambda = 0.054$ and remaining zero. These results highlight two key points with practical implications. First, even without allowing weight shifts, strong correlations among the control units mean that alternative weighting schemes can produce more conservative estimates, closer to zero than those obtained from the standard SC method. Second, only small deviations from the original weights are sufficient for the estimated effect to disappear. Together, these findings suggest that, given the high correlations among control units and the possibility of small hypothetical weight shifts, the impact of terrorism on the Basque Country’s GDP may be smaller compared with what the SC method would indicate, or even null.

Our inference results point to a similar conclusion. For all values of the weight-shift parameter λ , the 95% CI from our method contains zero, and the intervals become shorter as λ increases. Consequently, we cannot reject the null hypothesis of no effect on the Basque Country’s per capita GDP. When weight shifts are allowed, the evidence likewise shows no clear positive or negative effect, and the decreasing length of the CI as λ increases reflects greater precision and strengthens confidence in the absence of an effect. However, our confi-

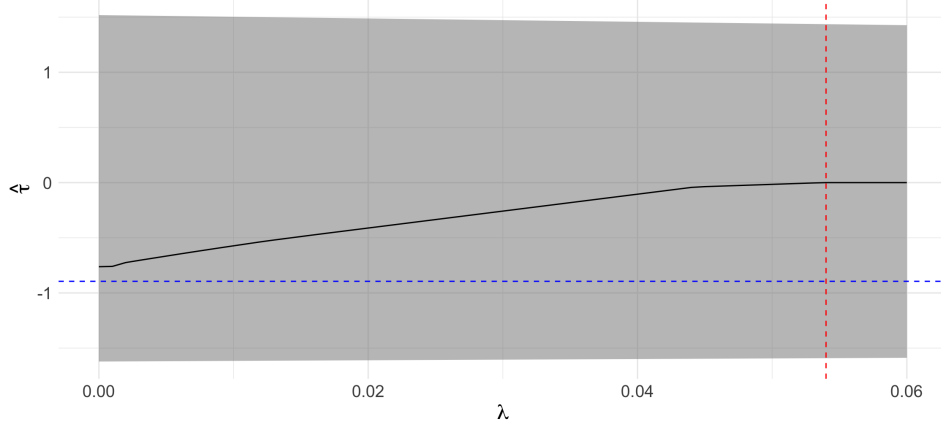


Figure 11: Reanalysis of the Basque study. The black solid line shows point estimates $\hat{\tau}$ in (18) and the gray area represents 95% CI $CI_{0.05}$ in (31) for each λ on x -axis; $\hat{\tau}^{SC}$ is shown as a blue dashed line, and the red line marks 0.054, where $\hat{\tau}$ first reaches 0.

dence interval tends to be slightly longer. One possible reason is the high correlation among control units, as shown in Figure 1. Such correlation inflates the covariance matrix in (23), leading to more dispersed perturbations and consequently wider CIs when aggregated as in (31).

9 Conclusion and Discussion

The SC method is a widely used tool for evaluating treatment effects in comparative case studies. Its validity, however, rests on critical assumptions: the control units are not highly correlated so that the weight vector used to construct the synthetic control is well-defined, and the linear relationship between the treated unit’s potential outcomes and those of the control units remains unchanged after treatment. To relax these conditions, we introduce a causal estimand derived from a DRO framework. This estimand coincides with the true time-averaged treatment effect when SC is identifiable, and otherwise serves as a conservative proxy. We further propose the DRoSC estimator, establish its convergence rate, and propose a novel perturbation inference tools, enabling principled inference even when the DRoSC estimator does not have a standard limiting distribution. This work generalizes SC, connects causal inference to DRO theory, and offers a foundation for robust inference in comparative case studies.

Several avenues for future research remain. A natural next step is to extend the framework to incorporate covariates, as originally proposed by Abadie and Gardeazabal (2003), Abadie et al. (2010, 2015). While some applications of the SC method use only outcomes, covariates can also improve the construction of counterfactuals for the treated unit (Abadie and Vivesi-Bastida, 2022). In our framework, this can be achieved by incorporating covariates into the uncertainty class, so that the weights reflect balance in both outcomes and covariates of the control units. Another direction is to relax the assumption of correctly specified weights within the simplex. Currently, we assume correct weight specification, but this may not hold in practice. While the best linear predictor can be considered within the simplex, its residuals are generally correlated with the predictor in a misspecified model. Consequently, our proposed uncertainty class in (9), which is constructed based on the covariance structure between these

two terms, may no longer be valid in this setting. Instead, our method can be generalized by redefining the uncertainty class, for example in terms of mean squared error (e.g., Xiong et al., 2023). Finally, extending the framework to settings with staggered adoption, where units receive treatment at different times, is both practically relevant and technically challenging (e.g., Athey and Imbens, 2022; Ben-Michael et al., 2022; Cattaneo et al., 2025). In such settings, the same identification challenges persist, stemming from high correlations among control units and weight shifts. While we define a new causal estimand, the weight-robust treatment effect, to generalize the time-averaged ATT under DRO framework, our method may be extended to staggered adoption by defining alternative reward functions corresponding to estimands beyond the time-averaged ATT.

References

- Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature*, 59(2):391–425.
- Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california’s tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.
- Abadie, A., Diamond, A., and Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. *Journal of Statistical Software*, 42:1–17.
- Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.
- Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132.
- Abadie, A. and L’hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817–1834.
- Abadie, A. and Vives-i-Bastida, J. (2022). Synthetic controls in action. *arXiv preprint arXiv:2203.06279*.
- Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51.
- Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pages 817–858.
- Andrews, D. W. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6):1341–1383.
- Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, pages 399–405.
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.
- Armstrong, T. B., Kolesár, M., and Kwon, S. (2023). Bias-aware inference in regularized regression models. *arXiv preprint arXiv:2012.14823v3*.

- 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*, 116(536):1716–1730.
- Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32.
- Athey, S. and Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79.
- Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.
- Ben-Michael, E., Feller, A., and Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536):1789–1803.
- Ben-Michael, E., Feller, A., and Rothstein, J. (2022). Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):351–381.
- Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.
- Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.
- Blackwell, D. A. and Girshick, M. A. (1979). *Theory of games and statistical decisions*. Courier Corporation.
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Cattaneo, M. D., Feng, Y., Palomba, F., and Titiunik, R. (2025). Uncertainty quantification in synthetic controls with staggered treatment adoption. *Review of Economics and Statistics*, pages 1–46.
- Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880.
- Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864.
- Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2025). Debiasing and t -tests for synthetic control inference on average causal effects. *arXiv preprint arXiv:1812.10820*.
- Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.
- Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979 – 1012.
- Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

- Ferguson, B. and Ross, B. (2020). Assessing the sensitivity of synthetic control treatment effect estimates to misspecification error. *arXiv preprint arXiv:2012.15367*.
- Firpo, S. and Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference*, 6(2):20160026.
- Fry, J. (2024). A method of moments approach to asymptotically unbiased synthetic controls. *Journal of Econometrics*, 244(1):105846.
- Gunsilius, F. F. (2023). Distributional synthetic controls. *Econometrica*, 91(3):1105–1117.
- Guo, Z. (2023a). Causal inference with invalid instruments: post-selection problems and a solution using searching and sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):959–985.
- Guo, Z. (2023b). Statistical inference for maximin effects: Identifying stable associations across multiple studies. *Journal of the American Statistical Association*, pages 1–17.
- Guo, Z., Li, X., Han, L., and Cai, T. (2025a). Robust inference for federated meta-learning. *Journal of the American Statistical Association*, pages 1–16.
- Guo, Z., Wang, Z., Hu, Y., and Bach, F. (2025b). Statistical inference for conditional group distributionally robust optimization with cross-entropy loss. *arXiv preprint arXiv:2507.09905*.
- Hahn, J. and Shi, R. (2017). Synthetic control and inference. *Econometrics*, 5(4):52.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
- Kuchibhotla, A. K., Balakrishnan, S., and Wasserman, L. (2024). The hulk: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):586–622.
- Li, K. T. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083.
- Liu, J., Tchetgen Tchetgen, E., and Varjão, C. (2024). Proximal causal inference for synthetic control with surrogates. In *International Conference on Artificial Intelligence and Statistics*, pages 730–738. PMLR.
- Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.
- Park, C. and Tchetgen Tchetgen, E. J. (2025). Single proxy synthetic control. *Journal of Causal Inference*, 13(1):20230079.
- Rosenbaum, P. R. (2002). *Observational Studies*. Springer New York.
- Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

- Shen, D., Ding, P., Sekhon, J., and Yu, B. (2023). Same root different leaves: Time series and cross-sectional methods in panel data. *Econometrica*, 91(6):2125–2154.
- Shi, X., Li, K., Miao, W., Hu, M., and Tchetgen Tchetgen, E. (2021). Theory for identification and inference with synthetic controls: a proximal causal inference framework. *arXiv preprint arXiv:2108.13935*.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.
- Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.
- Xie, M. and Wang, P. (2024). Repro samples method for a performance guaranteed inference in general and irregular inference problems. *arXiv preprint arXiv:2402.15004*.
- Xiong, X., Guo, Z., and Cai, T. (2023). Distributionally robust transfer learning. *arXiv preprint arXiv:2309.06534*.
- Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.
- Zeitler, J., Vlontzos, A., and Gilligan-Lee, C. M. (2023). Non-parametric identifiability and sensitivity analysis of synthetic control models. In *Conference on Causal Learning and Reasoning*, pages 850–865. PMLR.