Covariate Selection for the Synthetic Control Method

Byoung G. Park* Yongki Yoon†

January. 2026

Abstract

This paper proposes a new procedure to select appropriate covariates for the synthetic control method (SCM) using group square-root LASSO. Traditional SCM assumes that a set of time-invariant covariates share the same synthetic control weights as the outcome of interest. We consider time-varying covariates, some of which have different synthetic control weights from the outcome of interest. Our method consists of two-step procedure: in the first-step, we select only valid covariates using group square-root LASSO, and in the second-step we conduct the standard SCM, inspired by Post-LASSO approach. By re-estimating the synthetic weights using only the selected covariates, we can mitigate potential biases arising from mis-specified covariates. Moreover, our proposed estimator converges at least the same rate as the usual LASSO regardless of the variable selction. We evaluate the performance of the proposed method through Monte Carlo simulations and an empirical application.

Keywords: synthetic control method, variable selection, square-root LASSO, post-LASSO.

JEL Codes: C21, C23, C54

*bpark2@albany.edu
†yyoon@albany.edu

1 Introduction

Counterfactual analysis is widely used approach to estimate the causal impact of a particular intervention or treatment. The potential outcomes framework, pioneered by Rubin (1974), has been the dominant paradigm for the causal inference (see, Imbens and Rubin (2015) for more detailed discussion). Within this framework, Abadie and Gardeazabal (2003) and Abadie et al. (2010) propose a synthetic control method (SCM) to estimate the treatment effect of the treated unit(s) and later this method is elaborated in Abadie et al. (2015) (see, e.g., Abadie (2021) for detailed review). This approach has emerged as an important tool for estimating the treatment effect. Athey and Imbens (2017) assess that "the synthetic control method is arguably the most important innovation in the evaluation literature in the last fifteen years."

Since these seminal contributions, SCM has been widely adopted across various fields (see, e.g. Abadie et al., 2015; Ellison et al., 2024; Funke et al., 2023, or more). The versatility of the method has also led to the development of several extensions and variations, such as the synthetic control method with staggered adoption proposed by Ben-Michael et al. (2022), and the synthetic difference-in-differences method presented by Arkhangelsky et al. (2021). Some research studied statistical inference or properties of SCM (see, e.g. Chernozhukov et al. (2021); Ferman and Pinto (2021); Li (2020)). Other papers proposed novel approaches to combine synthetic controls with machine learnings such as LASSO (Least Absolute Shrinkage and Selection Operator) or Ridge regression (see, e.g., Abadie and L'Hour (2021); Ben-Michael et al. (2021); Kellogg et al. (2021)).

For clarity and depth of discussion, we begin by establishing the formal framework. Suppose we observe outcome variables Y_{jt} for $j = 1, \ldots, J$ over $t = 1, \ldots, T$ periods. Without loss of generality, we assume that only the first unit is treated at period $t > T_0$ where T_0 is the number of pre-treatment period. Following potential outcome framework, observed outcome

Y_{jt} can be divided by potential outcomes,

$$Y_{jt} = D_{jt}Y_{jt}(1) + (1 - D_{jt})Y_{jt}(0), \quad \text{for } j = 1, \ldots, J, t = 1, \ldots, T.$$

Here, D_{jt} is a treatment indicator such that $D_{jt} = 1$ if unit j is treated at time t and 0 otherwise. $Y_{jt}(1)$ is the potential outcome for unit j at time t with treatment and $Y_{jt}(0)$ is the potential outcome for unit j at time t without treatment. Our primary interest is to estimate the treatment effect, $\tau_{1t} = Y_{1t}(1) - Y_{1t}(0)$ at $t > T_0$. The challenge, of course, is that we cannot observe both potential outcomes after the treatment, so we have to estimate $Y_{1t}(0)$ for post-treatment period. Synthetic control method addresses this issue by constructing a synthetic control unit that best mimics the treated unit's counterfactual outcome, $Y_{1t}(0)$—a weighted average of donor pools that serves as a counterfactual for the treated unit. The synthetic control is created by forming a linear combination of control units with synthetic weight w^* that closely mirrors the characteristics of the treated unit in the pre-intervention period. It provides an estimate of the treated unit's trajectory had the intervention not occurred. So the method focuses on how to estimate the synthetic weight.

Here, we restrict the model to the linear factor model where the model is commonly discussed in most of the synthetic control method literature (see, e.g. Abadie (2021); Abadie et al. (2010)) such that:

$$Y_{jt}(0) = a_t + b_t X_j + \phi_t m_j + \epsilon_{jt}, \tag{1}$$

where a_t is a time trend, $X_j = (X_j^1, \ldots, X_j^q)$ is q-vector of observed covariates with coefficients b_t, m_j is unobserved covariates(i.e. factor loadings) with ϕ_t, and ϵ_{jt} follows white noise, i.e. $\epsilon_{jt} \sim WN(0, \sigma)$. Suppose there exists $(J-1) \times 1$ vector of weights $w^* = (w_2^*, \ldots, w_J^*)^\top$ such that

$$\sum_{j=2}^{J} w_j^* Y_{jt} = Y_{1t}, \quad \sum_{j=2}^{J} w_j^* X_j = X_1 \quad \text{for} \quad t = 1, \ldots, T_0. \tag{2}$$

If $m_1 = \sum_{j=2}^{J} w_j^* m_j$, then a synthetic control estimator yields an unbiased estimator of the treatment effect, $\widehat{\tau}_{1t} = Y_{1t}(1) - \widehat{Y}_{1t}(0)$. However, it is infeasible since m_j is unobserved. As highlighted in Abadie et al. (2010), a match in observed characteristics does not automatically guarantee a match in unobserved factors. If the chosen covariates are structurally irrelevant to the outcome, any observed match during the pre-treatment period may merely be a coincidence driven by the compensation of transitory shocks, ϵ_{jt}. Also, Abadie and Vives-i Bastida (2022) highlight the importance to include important covariates in the estimation as it allows lower interpolation bias, i.e. covariates excluded from X_j are automatically absorbed into m_j and it increases the bound of bias (see, e.g. Abadie (2021); Abadie and Vives-i Bastida (2022)). Meanwhile, including too many covariates may lead to overfitting, which may also increase the bound of bias.

Note that the synthetic control estimator is based on the assumption that outcome variable and covariates share the same weight structure, w^*. Hence, if a covariate has significantly different weight from outcome vairable, including the covariate may lead to bias in both estimation of weight and treatment effect. Although Botosaru and Ferman (2019) show that we can obtain such bounds of bias of estimated treatment effect in the absence of the assumption, we have to rely on stronger assumption such that the effects of observed and unobserved covariates on outcome of interest in the pre-treatment period should be linearly independent, and we have expense of deriving looser bounds. Hence, still the assumption on the weight structure is a key assumption of the synthetic control method. Then questions may arise; how can we select only the *valid* covariates that satisfy the assumption? Or, how can we detect whether the weight of a covariate is the same as that of outcome? For the sake of clarity, we shall use the terms *important* and *valid* interchangeably in this paper. *Important* covariates means those covariates which satisfies the assumption that they have same weight with outcome variables. Similarly, *unimportant* will be considered synonymous with *invalid* in terms of violating the assumption. Also, following Abadie and Vives-i Bastida (2022), we use the term *covariates* for X_{jt} only. *Predictors* often refers to

$(Y_{j1}, \ldots, Y_{jT_0}, X_j^1, \ldots, X_j^k)$, the combination of pre-treatment outcome and a set of covariates which is used for estimating the synthetic control weight. Later, we specifically distinguish our work from Vives-i Bastida (2022), who focuses on the selection of *predictors* rather than *covariates*, necessitating a clear conceptual differentiation between the two.

However, the standard synthetic control method does provide little guidance about how to select *important* covariates given their information as well as the practitioners have little idea about which covariates are *important*, while some previous studies had highlighted the importance of this issue. (see, e.g. Abadie (2021); Abadie et al. (2015); Ferman et al. (2020); Botosaru and Ferman (2019)) Recently, Vives-i Bastida (2022) explored this issue and propose a sparse synthetic control method by formulating a bi-level optimization problem by imposing an penalty, thereby selecting only the important predictors. Vives-i Bastida (2022) show that the sparse synthetic control method achieves lower bias and has better post-treatment performance than the standard synthetic control method.

While we follow this line of research, our study differs in both the target aned structure of the selection procedure. We employ group square-root LASSO, proposed by Bunea et al. (2014), an extension of square-root LASSO by Belloni et al. (2011), for variable selection. Unlike previous SCM literature, we allow covariates to be time-varying, thereby allowing for a more flexible modeling framework to directly penalize the difference between the outcome weight and covariate-specific weight. This approach enables us not only to estimate the synthetic weights but also to identify the *valid* covariates. We divide our estimation procedure into two-stage: in the first stage, we select only the *valid* covariates using group square-root LASSO, and in the second stage, we conduct the standard synthetic control method using only the selected covariates, inspired by Post-LASSO approach (see, Belloni and Chernozhukov (2013)). Our proposed two-step method has several advantages. First, by re-estimating the synthetic weights using only the selected covariates, we can mitigate potential biases arising from mis-specified covariates. Second, our method inherits the oracle property from LASSO, ensuring consistency in covariate selection and estimation under

suitable conditions.

The rest of the paper is organized as follows. In Section 2, we introduce our model and estimation procedure. In Section 2.1, we discuss the basic idea of our method. In Section 2.2, we introduce the first-stage estimation using group square-root LASSO, and in Section 2.3, we discuss the second-stage estimation using the standard synthetic control method. In Section 2.4 we discuss the choice of tuning paramter. In Section 3, we present the results of Monte Carlo simulations to evaluate the finite-sample performance of our method. Finally, Section 4 concludes.

2 Models

2.1 The settings

As discussed in Section 1, we consider a linear factor model as Equation (1). The standard way to estimate the synthetic control weight is the least squares using the pretreatment period observations. As $Y_{jt}(0)$ is observed for all $j = 2, \ldots, J$ when $t = 1, \ldots, T_0$, one can estimate the weights that minimize an empirical distance between Y_{1t} and $\sum_{j=2}^{J} \hat{w}_j Y_{jt}$ provided that the pretreatment period is long enough.

In practice, some covariates are often used to estimate the synthetic weight. Assume that there exists time-constant covariates X_j^k for $k = 1, \ldots, q$ that satisfies following condition.

Condition 1a (Perfect weight matching on covariates)**.** *There exists a set of weights* $w^* = (w_2^*, \cdots, w_J^*)^\top$ *such that*

$$X_1^k = \sum_{j=2}^{J} w_j^* X_j^k, \quad \textit{for } k = 1, \cdots, q. \tag{3}$$

where w_j^* *is the same weight for outcome variable, i.e.* $Y_{1t}(0) = \sum_{j=2}^{J} w_j^* Y_{jt}(0)$.

Condition 1a ensures that the covariates have the same weight as the outcome variable. Including these covariates may improve the estimation of treatment effects while allow some

mis-fitting in pre-treatment periods.

Let $\tilde{Z}_1$ and $\tilde{Z}_0$ be matrices such that $\tilde{Z}_1 = (\mathcal{Y}_1, X_1^1, \cdots, X_1^q)^\top$ and $\tilde{Z}_0$ has same pre-treatment characteristics as Z_1 but for controlled units. Here, $\mathcal{Y}_1$ denotes the stacked outcome vectors for treated unit, and $\mathcal{Y}_0$ denotes that for controlled units. We do not have to include all lagged outcome variables in $\tilde{Z}_1$ and $\tilde{Z}_0$. Moreover, previous studies allow X_j^k to be able to have some linear combination of outcome of interest. For example, Abadie et al. (2015) used the average of outcome of interest, GDP per capita for the 1981-90 period as a covariate. However, in later discussion, we allow covariates to vary over time, thus we exclude such case where average of outcome of interest is seen as a covariate.

We can estimate the weight vector by minimizing the following optimization problem:

$$\tilde{w} = \arg\min_{w} \|\tilde{Z}_1 - \tilde{Z}_0 w\|_V^2, \quad \text{subject to } w_j \geq 0 \text{ and } \sum_{j=2}^{J} w_j = 1, \tag{4}$$

where $\|A\|_V = \sqrt{A^\top V A}$ is weighted norm with positive semi-definite diagnoal matrix V. This matrix V reflects relative importance of variables in Z_1 and Z_0 (see, e.g., Abadie et al. (2010); Abadie et al. (2015) for more details). For example, the matrix V may represent the time-dimension importance of outcome variable or also may indicate the relative importance among the covariates, since their *predictors* include both covariates and outcome variables. However, this characterization conflates several distinct roles that V can play. We distinguish three conceptually separate roles that V can fulfill. First, the diagonal elements of V can serve to normalize variables measured on different scales. However, this role becomes redundant when the data are standardized prior to estimation. Since standardization is a common preprocessing step in synthetic control applications, this role of V is not fundamental to the methodology. Second, for the outcome variable, V can assign differential weights across time periods. In a setting with T_0 pre-treatment periods, the first T_0 diagonal elements of V determine the relative importance of each time period in constructing the synthetic control. This is particularly relevant when practitioners believe that more recent pre-treatment obser-

vations are more informative about the counterfactual trajectory. Last, when time-invariant covariates are included as predictors, V also governs the relative importance among different covariates. In the traditional synthetic control method framework where predictors form a $(T_0 + q) \times 1$ vector, a diagonal V matrix of dimension $(T_0 + q) \times (T_0 + q)$ simultaneously encodes both timewise importance—through its first T_0 elements—and covariate importance—through its last q elements. This conflation makes it difficult to separately identify and interpret these two distinct concepts.

Our approach achieves a clear separation of these roles. By allowing covariates to be time-varying and employing the group LASSO, the task of covariate selection is delegated to the penalty structure rather than the weight matrix. This allows V to focus solely on timewise importance for the outcome variable, while covariate importance is handled through the group LASSO penalty. This separation offers two advantages. First, it provides conceptual clarity: the penalty addresses the question of *which* covariates to include, while V addresses the question of *how much* to weight each time period. Second, it aligns naturally with our framework where a timewise weighting scheme can be applied uniformly across all variables without confounding the selection mechanism.

The primary contribution of this study lies in the development of a two-step approach for covariate selection. Building upon previous research, we employ the group square-root LASSO method introduced by Bunea et al. (2014) for variable selection in synthetic control method. The group square-root LASSO is an extension of square-root LASSO by Belloni et al. (2011) and group LASSO by Yuan and Lin (2006).

Before introducing our method, we first go back to the questions that we raised in the introduction. The question that we want to address is how we can detect whether a covariate satisfies the condition 1a, i.e. a covariate is *important*. To answer this question, suppose that we have k-th covariate X^k has weight as $\psi^k = w^* + \delta^k$ for $k = 1, \ldots, q$ where δ^k is the difference between two weights. Then, if δ^k is zero, we can assure that the k-th covariate satisfies 1a and is *valid*. On the other hand, if δ^k is far from zero, the k-th covariate is

unimportant. Thus, we aim to select only the valid covariates that satisfies the condition 1a. To achieve this goal, we allow the covariates to be time-varying. This approach gives us more flexibility in selecting the covariates by directly penalizing the difference between two weights, δ^k. We exclude cases in which the average value of the outcome of interest was treated as one of the covariates. In this approach, we have modified version of Condition 1a as follows:

Condition 1b (Perfect weight matching on time-varying covariates)**.** *There exists a set of weights* $w^* = (w_2^*, \cdots, w_J^*)^\top$ *such that*

$$X_{1t}^k = \sum_{j=2}^{J} w_j^* X_{jt}^k, \quad \textit{for } k = 1, \cdots, q,\ t = 1, \cdots, T_0. \tag{5}$$

where w_j^* *is the same weight for outcome variable, i.e.* $Y_{1t}(0) = \sum_{j=2}^{J} w_j^* Y_{jt}(0)$.

Based on our settings, we have

$$X_{1t}^k = \sum_{j=2}^{J} \psi_j^k X_{jt}^k \equiv \sum_{j=2}^{J} (w_j^* + \delta_j^k) X_{jt}^k, \quad \text{for } k = 1, \cdots, q,\ t = 1, \cdots, T_0. \tag{6}$$

Following this structure, we can construct design matrices Z_1 and Z_0 such that

$$Z_1 = (\mathcal{Y}_1, \mathcal{X}_1^1, \ldots, \mathcal{X}_1^q)^\top \tag{7}$$

$$Z_0 = \begin{pmatrix} \mathcal{Y}_0 & 0 & \cdots & 0 \\ \mathcal{X}_0^1 & \mathcal{X}_0^1 & 0 & 0 \\ \vdots & \cdots & \ddots & \vdots \\ \mathcal{X}_0^q & 0 & \cdots & \mathcal{X}_0^q \end{pmatrix}, \tag{8}$$

where $\mathcal{X}_1^k$ is the stacked vector of k-th covariate for treated unit, and $\mathcal{X}_0^k$ is the stacked matrix of k-th covariate for controlled units. Here, Z_1 is $(1+q)T_0 \times 1$ vector and Z_0 is $(1+q)T_0 \times (1+q)(J-1)$ matrix. For example, if we have only two covariates i.e., $q = 2$.

Then, Z_1 and Z_0 becomes

$$Z_1 = (\mathcal{Y}_1, \mathcal{X}_1^1, \mathcal{X}_1^2)^\top$$
$$Z_0 = \begin{pmatrix} \mathcal{Y}_0 & 0 & 0 \\ \mathcal{X}_0^1 & \mathcal{X}_0^1 & 0 \\ \mathcal{X}_0^2 & 0 & \mathcal{X}_0^2 \end{pmatrix}.$$

The difference between $\tilde{Z}_1$ in traditional synthetic control and Z_1 is that the design matrix Z_1 contains all timewise information not only of outcome variable but also covariates while $\tilde{Z}_1$ have individual information of covariates.

We also define the group-structured weight matrix ω such that

$$\omega = \begin{pmatrix} w^* \\ \delta^1 \\ \vdots \\ \delta^q \end{pmatrix}, \tag{9}$$

Then, we can solve the following optimization problem to obtain the optimal weight matrix $\widehat{\omega}$ such that:

$$\widehat{\omega} = \arg\min_{\omega} \left(\frac{1}{\sqrt{n}} \| Z_1 - Z_0 \omega \|_2 + \frac{\lambda}{n} \sum_{k=1}^{q} \tilde{\omega}^k \sqrt{p^k} \| \delta^k \|_2 \right), \tag{10}$$

where $n = (1+q) * T_0$ is the number of rows for Z_0, λ is the penalty parameter, $\tilde{\omega}^k$ is an initial groupwise estimate—groupwise norm of estimate—of ω^k obtained from the unpenalized model, and p^k accounts for size of group. We can usually use OLS estimator for $\tilde{\omega}^k$ but if it is doubted multicollinearity, Ridge regression estimator is used following Yuan and Lin (2006). Since Yuan and Lin (2006) or other group LASSO scheme assume that the group is pre-defined, we suppose that the number of groups K is known and defined as $K = q + 1$ and for each group, the size is known as $p^k = p = J - 1$.

This optimization form is similar to the group square-root LASSO studied by Bunea et al. (2014), which is an extension of square-root LASSO by Belloni et al. (2011) and group LASSO by Yuan and Lin (2006). The first term in the objective function represents the square-root loss function, while common LASSO employs the squared loss function. The second term is the group LASSO penalty which penalizes the group-level differences between outcome weight and covariate-specific weights. By minimizing this objective function, we can obtain correct pattern recovery of which covariates are *valid* which satisfy Assumption **??** unders suitable conditions, with a σ-independent penalty level.

We can obtain the estimated weight $\widehat{\omega}$ and they seem like:

$$\widehat{\omega} = \begin{pmatrix} \widehat{w} \\ \widehat{\delta}^1 \\ \vdots \\ \widehat{\delta}^q \end{pmatrix}.$$

Thus first group of $\widehat{\omega}$, which is $\widehat{w}$ is the estimated synthetic weight, which is our primary interest. The other groups, $\widehat{\delta}^k$ for $k = 1, \ldots, q$ are the estimated difference between $\widehat{w}$ and $\widehat{\psi}^k$, and they gives the information about which covariates are useful. If $\widehat{\delta}^k$ is estimated as zero, it means that the k-th covariate is useful to construct the synthetic weight, which implies that the covariate satisfies the assumption that its weight is same as outcome variable's weight, w^*. On the other hand, if $\widehat{\delta}^k$ is estimated as non-zero, it means that the k-th covariate is not valid to construct the synthetic weight. With estimated weight vector, the treatment effect of the treated at time $t > T_0$ is estimated by

$$\widehat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J} \widehat{\omega}_j Y_{jt}(0). \tag{11}$$

2.2 First-stage: Variable Selection with Group Square-root LASSO

Our proposed method consists of a two-stage estimation proceudre: in the first stage, we employ the Group Square-root LASSO to identify and select the valid covariates that satisfy Assumption 1b. Square-root LASSO is a pivotal method proposed by Belloni et al. (2011) in that it does not require knowledge of the standard deviation of the error term. It was further extended to the group structure by Bunea et al. (2014). Group square-root LASSO uses square-root loss function combined with group LASSO penalty, having the same accuracy as the group LASSO but with a σ-independent penalty level.

Unlike the conventional LASSO framework, where *important* covariates are identified by non-zero coefficients, our approach is designed such that the difference between outcome and covariate weights, δ is penalized. Consequently, the valid covariates—those possessing identical synthetic control weights as the outcome variable—are identified when their corresponding group-level differences are shrunk to exactly zero. In the second stage, we implement a Post-LASSO approach, proposed by Belloni and Chernozhukov (2013), by applying the standard synthetic control method to the selected subset of covariates. This refinement step effectively mitigates the shrinkage bias inherent in penalized estimators and allows us to obtain theoretical properties, inclduing an improved convergence rate. We will discuss this in more detail in Section 2.3.

Before discussing the proposed method, we first make some assumptions. Define the set of bias terms, $S = \{k : \delta^k \neq 0\} \cup \{0\}$. Zero index in S represents the outcome variable group. Let $\mathcal{S} = S \setminus \{0\}$ be a set of invalid covariates, then $S^C = \mathcal{S}^C$, the inactive set, is a set of valid covariates since our analysis take value 0 for valid covariates. Also, let $\dot{s} = |\mathcal{S}|$ be the cardinality of the set S and we assume that $\dot{s}$ is small compared to the total number of groups $K = q + 1$.

Assumption 1 (Group Irrepresentable Condition)**.** *There exists* $\eta \in (0, 1)$ *such that*

$$\max_{k \notin S} \|\mathcal{X}_0^{k\top} Z_0^{(S)} (Z_0^{(S)\top} Z_0^{(S)})^{-1} u^{(S)}\|_2 < 1 - \eta, \tag{12}$$

, where $Z_0^{(S)}$ is the submatrix of Z_0 corresponding to the support set S, and $u^{(S)}$ is a subgradient vector for all $k \in S$.

This assumption is a necessary and sufficient condition for consistent support recovery via LASSO-type methods (see, e.g. Zhao and Yu (2006); Bunea et al. (2014)). It ensures that the invalid groups of covariates are not too correlated with the valid groups, allowing for accurate selection of valid covariates.

Remark 1 (**Intuitive Interpretation under Orthogonality**). Assumption 1 can be difficult to be interpreted directly. To clarify its geometric meaning, consider a simplified scenario where the predictors in the active set S are mutually orthogonal.

Under the assumption of orthogonality, the Gram matrix $Z_0^{(S)\top} Z_0^{(S)}$ becomes a diagonal matrix with diagonal entries euqal to the squared norms of the variables, i.e. $Z_0^{(S)\top} Z_0^{(S)} \approx \text{diag}(\|\mathcal{Y}_0\|_2^2, \|\mathcal{X}_0^k\|_2^2, \dots)$. Consequently, its inverse matrix is also diagonal. Define v_s be s-th column of $Z_0^{(S)}$, then we have:

$$\mathcal{X}_0^k Z_0^{(S)} (Z_0^{(S)\top} Z_0^{(S)})^{-1} u^{(S)} = \left[\frac{< \mathcal{X}_0^k, v_1 >}{\|v_1\|_2^2}, \dots, \frac{< \mathcal{X}_0^k, v_s >}{\|v_s\|_2^2} \right]$$

where $< \cdot, \cdot >$ denotes the inner product. Thus, Assumption 1 reduces to:

$$\frac{\|\mathcal{X}_0^k\|_F^2}{\|\mathcal{Y}_0\|_2^2 + \sum_{s \in \mathcal{S}} \|\mathcal{X}_0^s\|_F^2} < 1 - \eta, \quad \text{for all } k \notin \mathcal{S}, \tag{13}$$

where $\|\cdot\|_F$ is the Frobenius norm. This inequality (13) implies that for consistent variable selection, the variation of any single valid covariate (numerator) must not dominate the total variation of the *active subspace* spanned by the outcome and the invalid covariates (denominator). In other words, if a valid covariate—which should have $\delta^k = 0$—exhibits excessively high variation or correlation relative to the signal from the invalid covariates, the LASSO might mistakenly select it to explain the residual variance. Thus, Assumption 1 essentially requires that valid covariates be sufficiently distinguishable from the linear combination of invalid covariates and the outcome. This condition may hold when we standardize outcome

and each covariate to have unit norm before estimation. $\Diamond$

Assumption 2 (Beta-min Condition)**.** *For any biased covariates $k \in \mathcal{S}$, the magnitude of bias $\|\delta^k\|_2$ must be large enough to be distinguished from noise. Formally, there exists a constant $D > 0$ such that*

$$\min_{k \in \mathcal{S}} \|\delta^k\|_\infty \geq D \frac{\sigma \lambda \tilde{\omega}^k \sqrt{p^k}}{n}. \tag{14}$$

Assumption 2 ensures that the true non-zero coefficients are sufficiently large compared to the noise level, allowing the group square-root LASSO to correctly identify the active groups of covariates.

Assumption 3 (Compatibility Condition)**.** *There exists a constant $\kappa > 0$ and $\gamma > 1$ such that for all $\Delta \in \Delta_\gamma$,*

$$\sum_{k \in S} \tilde{\omega}^k \sqrt{p^k} \|\Delta^k\|_2 \leq \frac{\sqrt{s^*} \|Z_0 \Delta\|_2}{\sqrt{n} \kappa}, \tag{15}$$

where $\Delta_\gamma = \{\Delta \in \mathbb{R}^q : \sum_{k \in S^C} \tilde{\omega}^k \sqrt{p^k} \|\Delta^k\|_2 \leq \gamma \sum_{k \in S} \tilde{\omega}^k \sqrt{p^k} \|\Delta^k\|_2\}$ and $s^* = \sum_{k \in S} p^k$ is the total size of groups in the set S. Here, the compatibility constant κ measures the correlation in the design matrix: the smaller the value of κ, the higher the correlation among variables.

While Assumption 3 is fundamental for establishing bounds on prediction and estimation errors, the consistent recovery of the true sparsity pattern typically requires the stronger assumption, Assumption 1. It is well-established that the Assumption 1 is strictly more restrictive than the Assumption 3; thus, if the design matrix satisfies the condition necessary for correct support recovery, it automatically satisfies the requirement for optimal prediction accuracy.

Assumption 4 (Sparsity Condition)**.** *The sparsity index s^* satisfies:*

$$s^* < \frac{n^2\kappa^2}{\lambda^2} \tag{16}$$

where $\kappa > 0$ is the compatibility constant defined in Assumption 3.

Bunea et al. (2014) show that $s^* \lesssim \frac{n}{\log q}$ and highlight that regardless of allowing the situation $n < p$, we cannot expect good performance of variable selection unless the true model has fewer parameters than n. We first study the model selection consistency of our proposed method. Following Bunea et al. (2014), define the noise component as:

$$V = \max_{1\le k\le q} \frac{\|\sqrt{n}(Z_0^{k\top}\varepsilon)^k\|_2}{\sqrt{p^k}\|\varepsilon\|_2}, \tag{17}$$

where $\varepsilon = (Z_1 - Z_0\omega^*)/\sigma$. For given $\lambda > 0$, let $\mathcal{A} = \{V \le \lambda/\bar{\gamma}\}$ where $\bar{\gamma} = \frac{\gamma+1}{\gamma-1}$. Next, define the event $\mathcal{A}_1 = \{V \le \lambda/(\bar{\gamma} \vee 2\bar{\eta})\}$ where $\bar{\eta} = \frac{1+\eta}{1-\eta}$ and η is defined in Assumption 1.

Theorem 1 (Model Selection Consistency)**.** *Assume that Assumptions 1, 2, 3 and 4hold. Let $D > 0$ be a dominating constant, then on the set $\mathcal{A}_1$, with probability at least $1-\alpha$, we have*

1. $\widehat{\omega}^{S^C} = 0$;
2. *For all $1 \le g \le q$, $\|(\widehat{\omega} - \omega^*)^g\|_\infty \le D\frac{\lambda\sigma\tilde{\omega}^g\sqrt{p^g}}{n}$;*
3. *If there exists an m such that $m_g \ge D\frac{\lambda\sigma\tilde{\omega}^g\sqrt{p^g}}{n}$ for all $g \in \mathcal{S}$, then $\mathcal{S} = \widehat{\mathcal{S}}$.*

See the proof in Appendix A.

Remark 2. The proof proceeds without explicitly accounting for the constraints on the weight vector, i.e., non-negativity and summing to one. This is justified because: (i) the non-negativity constraint introduces additional terms in the KKT conditions, which only strengthens our contradiction argument; (ii) the sum-to-one constraint introduces a term in-

volving that can be shown to be of smaller order $O(\lambda/n)$ and does not affect the leading-order analysis. $\Diamond$

Constant D depends on κ, γ, η, and $\zeta_{\|\cdot\|_\infty}$, but not on n, p^k, or q. Theorem 1 establishes that under suitable conditions, the group square-root LASSO can consistently identify the valid covariates whose weights match those of the outcome variable. Specifically, part 1 shows that the estimated coefficients for valid covariates are shrunk to exactly zero. Part 2 provides a bound on the estimation error for the coefficients of invalid covariates, while part 3 guarantees that if the true non-zero coefficients, $\widehat{\delta}^k$ for $k \in \mathcal{S}$, are sufficiently large compared to the noise level, the method will correctly recover the active set of covariates. This result is crucial for ensuring that our two-stage procedure effectively selects the appropriate covariates for constructing the synthetic control weights.

2.3 Second-stage: Post-LASSO Approach

After selecting the valid covariates using the group square-root LASSO in the first step, we proceed to the second step where we apply the standard synthetic control method to the selected subset of covariates. This Post-LASSO approach, inspired by Belloni and Chernozhukov (2013), aims to mitigate the shrinkage bias introduced by the penalized estimation in the first stage. By re-estimating the synthetic control weights using only the covariates identified as valid in the first step, we can obtain more accurate estimates of the treatment effects.

Formally, let $\widehat{\mathcal{S}}^C$ be the set of selected valid covariates from the first-stage, i.e. $\widehat{\mathcal{S}} = \{k : \|\widehat{\delta}^k\|_2 > 0\}$. We then construct new design matrices $Z_1^{\text{Post}} = (\mathcal{Y}_1, \mathcal{X}_1^1, \ldots, \mathcal{X}_1^{\widehat{q}})^\top$ and $Z_0^{\text{Post}} = (\mathcal{Y}_0^\top, \mathcal{X}_0^{1\top} \ldots, \mathcal{X}_0^{\widehat{q}\top})^\top$ where $\widehat{q} = |\widehat{\mathcal{S}}^C|$ is the number of valid covariates distinguished in the first-stage. The Post-LASSO estimator of the synthetic control weights, denoted as

$\widehat{w}^{\text{Post}}$, is obtained by solving the following optimization problem:

$$\widehat{w}^{\text{Post}} = \arg\min_{w} \|Z_1^{\text{Post}} - Z_0^{\text{Post}} w\|_2^2, \quad \text{subject to } w_j \geq 0 \text{ and } \sum_{j=2}^{J} w_j = 1. \tag{18}$$

This re-estimation step allows us to refine the synthetic weights. By focusing solely on the covariates deemed valid in the first stage, we can reduce the bias associated with the initial penalized estimates. The resulting Post-LASSO estimator $\widehat{w}^{\text{Post}}$ is expected to exhibit improved statistical properties, including a faster convergence rate and reduced mean squared error.

Assumption 5 (Restricted Eigenvalue Condition)**.** *For the design matrix Z_0 and a constant $\gamma > 1$, following condition holds:*

$$\bar{K}(\bar{\gamma}) = \min_{\Delta \in \Delta_{\bar{\gamma}}} \frac{\|Z_0 \Delta\|_2}{\sqrt{n}\|\Delta^{\mathcal{S}}\|_2} > 0, \tag{19}$$

where $\bar{\gamma} = (\gamma+1)/(\gamma-1) > 0$ and $\Delta_{\bar{\gamma}}$ is similarly defined as Δ_γ in Assumption 3.

This condition is commonly used and known to be plausible and general (see, e.g., Belloni and Chernozhukov (2013)). It ensures that the design matrix Z_0 is well-behaved in the subspace spanned by the active covariates, preventing issues of multicollinearity that could adversely affect the estimation accuracy of the Post-LASSO estimator. Now we establish the prediction error bound for the Post-LASSO estimator $\widehat{w}^{\text{Post}}$ defined in (18) following Belloni and Chernozhukov (2013).

Theorem 2 (Prediction Error Bound for Second-Stage Estimator)**.** *Suppose that 5 holds. Assume that the noise $Z_1^{Post} - Z_0^{Post} w^{Post}$ follows Gaussian, where $\widehat{w}^{\text{Post}}$ denote the second-stage estimator defined in* (18)*. Then, there exists a constant $C_{\tilde{K}} > 0$ such that with probability greater than $1-\alpha$,*

$$\frac{1}{\sqrt{n^{Post}}} \|Z_0^{Post}(\widehat{w}^{Post} - w^*)\|_2 \leq 2\sigma^{Post} \left(\frac{\sqrt{J} + \sqrt{C_{\tilde{K}} \dot{s} \log q}}{\sqrt{n^{Post}}} \right) + \sqrt{R_n}, \tag{20}$$

where $\dot{s} = |\mathcal{S}|$ *is cardinality of the set of invalid covariates and* $R_n = \min_{w \in \mathbb{R}^{\widehat{q}}} \frac{1}{n} \|Z_1^{\text{Post}} - Z_0^{\text{Post}} w\|_2^2$ *is the approximation error when the true model is not exactly sparse.*

See the proof in Appendix B. Theorem 2 establishes a bound on the prediction error of the Post-LASSO estimator, demonstrating its improved accuracy compared to the initial estimates. The bound consists of two components: the first term reflects the estimation error, which increases with $\widehat{s}$ and decreases with sample size n, while the second term, R_n, captures the approximgation error that arises when the true model is not exactly sparse. Theorem 2 is particularly valuable since it does not require the assumption of perfect model selection. Even though the first-stage may not always identify the exact set of valid covariates, the second-stage estimator still achieves a favorable convergence rate, at least same as that of the initial group square-root LASSO estimator. Building on Theorem 2, we can further characterize the asymptotic behavior of the Post-LASSO estimator when the model selection is perfect.

Remark 3. Although Theorem 2 does not assume perfect model selection nor other assumptions that is necessary for Theorem 1, it does depend on the Gaussian noise assumption, which was not required in Theorem 1. ◊

Corollary 1 (Convergence rate for second-stage estimator)**.** *Suppose that the conditions of Theorem 2 hold. Let* $\widehat{w}^{Post}$ *be the Post-LASSO estimator defined in* (18)*, and* $\widehat{w}^{Oracle}$ *be the infeasible oracle estimator obtained by solving* (18) *with the true model* $\mathcal{S}^C$*. Then, we have:*

1. *As* $n \to \infty$*, the probability of selecting the exact true model approaches 1, i.e.,* $P(\widehat{\mathcal{S}}^C = \mathcal{S}^C) \to 1$ *or* $P(\widehat{\mathcal{S}} = \mathcal{S}) \to 1$*.*

2. *Conditional on the event of perfect selection,* $\{\widehat{\mathcal{S}}^C = \mathcal{S}^C\}$*, the second-stage estimator is numerically identical to the oracle estimator, i.e.,* $\widehat{w}^{\text{Post}} = \widehat{w}^{\text{Oracle}}$*.*

3. *Consequently, the second-stage estimator achieves the oracle convergence rate. With*

probability approaching 1, we have

$$\frac{1}{\sqrt{n^{Post}}}\|Z_0^{\text{Oracle}}(\widehat{w}^{\text{Post}} - w^*)\|_2 = O_p\left(\sigma^{Post}\sqrt{\frac{\dot{s}}{n^{Post}}}\right), \tag{21}$$

where Z_0^{Oracle} is the design matrix constructed using only the valid covariates in the oracle model $\mathcal{S}^C$.

See the proof in Appendix C.

Remark 4. Estimating weights under a convex constraint set creates a contraction of the error, as discussed in Theorem 1. Specifically, let w^{ols} be the unconstrained minimizer on the selected support and w^{Post} be the constrained minimizer. By the properties of projection onto a convex set, $\|\widehat{w}^{\text{Post}} - w^*\|_2 \leq \|w^{\text{ols}} - w^*\|_2$. Therefore, the bound derived for the unconstrained case serves as a valid, and potentially conservative, upper bound for our second-stage estimator under the convex constraints. This confirms that the convergence rate established in 1 holds even when the synthetic control weights are estimated within the convex set. ◊

Corollary 1 further elaborates on this result by stating that as the sample size increases, the second-stage estimator is numerically identical to the oracle estimator, achieving the optimal convergence rate of $O_p\left(\sigma\sqrt{\frac{\dot{s}}{n}}\right)$. This highlights the effectiveness of our two-stage procedure in accurately estimating synthetic control weights when valid covariates are correctly identified.

2.4 Tuning Parameter

An important aspect of implementing the group square-root LASSO is the selection of the tuning parameter λ. The choice of λ influences the sparsity of the solution. While common LASSO often relies on cross-validation for tuning parameter selection, square-root LASSO offers an alternative approach that does not require knowledge of the noise level σ. Following Belloni et al. (2011) and Bunea et al. (2014), we propose a theoretical-driven choice for λ.

Specifically, we suggest settings λ as:

$$\lambda = n\sqrt{\frac{\max(\|Z_0^k\|^2/n)F^{-1}_{J-1,n-J+1}((1-\alpha)/q)}{(J-1)F^{-1}_{J-1,n-J+1}((1-\alpha)/q)+n-J+1}}, \tag{22}$$

where $F^{-1}_{J-1,n-J+1}((1-\alpha)/q)$ is the $((1-\alpha)/q)$ quantile of the F-distribution with degrees of freedom $J-1$ and $n-J+1$. This choice of λ ensures that the event $\mathcal{A}_1$ occurs with probability at least $1-\alpha$, thereby satisfying the conditions required for model selection consistency as established in Theorem 1. By adopting this theoretically grounded approach to tuning parameter selection, we can effectively balance the trade-off between bias and variance in our estimates without relying on computationally intensive cross-validation procedures.

3 Simulation study

We conduct simulation studies to evaluate the finite sample performance of our proposed method. We compare our two-stage synthetic control method with group square-root LASSO against the standard synthetic control method and the pre-testing method. The pre-testing method involves also two steps: first, it runs a linear regression of Z_1 on Z_0 to test Condition 1b, i.e., $H_0 : \delta^k = 0$ for each covariate k, and then constructs the synthetic control using only the covariates that fail to reject the null hypothesis. However, standard likelihood ratio test with bootstrap methods are not directly applicable for inference when parameters lie on the boundary by inequality constraints, as discussed in Andrews (2000) and Andrews (2001). In our framework, although δ^k is not directly on the boundary of inequality constraints, w^* does lie on the boundary, $w^* \geq 0$. Moreover, the covariate weight should be non-negative in the alternative hypothesis, i.e., $\psi^k = w^* + \delta^k \geq 0$. Also, they should satisfy sum-to-one constraints. In this case, standard subsampling or bootstrap may not a suitable approach. We employ shrinkage bootstrap, proposed to solve the boundary issue by Cavaliere et al. (2022), to conduct the hypothesis testing in the pre-testing method. We assess the accuracy of synthetic weight estimation across the different scenarios, based on mean squared error

(MSE) and bias.

For simplicity, we generate data according to the following data generating process:

$$
\begin{aligned}
&Y_{jt}(0) \sim N(0, \sigma^2), \quad X_{jt}(0) \sim N(0, \Sigma), \quad \text{for } j = 2, \ldots, J,\ t = 1, \ldots, T_0,\ k = 1, \ldots, q, \\
&Y_{1t}(0) = \sum_{j=2}^{J} w_j^* Y_{jt}(0) + \varepsilon_t, \quad \text{for } t = 1, \ldots, T_0, \\
&X_{1t}^k(0) = \sum_{j=2}^{J} (w_j^* + \delta_j^k) X_{jt}^k(0) + \varepsilon_t, \quad \text{for } t = 1, \ldots, T_0,\ k = 1, \ldots, q,
\end{aligned}
$$

where Σ is variance-covariance matrix of covariates, which allows for correlation among covariates. We set $J = 11$, i.e., 1 treated unit with 10 controlled units, and $T_0 = 50$. The true synthetic weight vector is set as $w^* = (1/(J-1), \ldots, 1/(J-1))$ whereas $\delta^k = (0, 7, 0.1, -1/(J-1), \ldots, -1/(J-1))$ for $k \in \mathcal{S}$. Note that $\delta^k = 0$ for $k \in \mathcal{S}^C$. The noise level is set to $\sigma = 1$. We fix $q = 6$ where $\dot{s} = \dot{q} = 3$, i.e., three valid covariates and three invalid covariates.

First, we evaluate the performance of our method under different number of $(J-1)/T_0$ ratios, specifically, $J - 1 = 10, 25, 40$ while keeping $T_0 = 50$. The results are summarized in Table 1 and Table 2. Our proposed method consistently outperforms both the standard synthetic control method and the pre-testing method across all scenarios. Notably, as the number of control units increases, the performance gap between our method and the alternatives widens, highlighting the robustness of our approach in more poor degree of freedom settings. Note that the pre-testing method shows poor performance when the number of control units is large, which may be due to the low power of the hypothesis testing under high-dimensional settings.

Next, we assess the impact of varying the number of covariates q on the performance of our method. WE consider scenarios with $q = 2, 6, 10$ while keeping $J - 1 = 10$ and $T_0 = 50$. The results, presentted in Table 3 and Table 4, demonstrate that our method maintains superior performance compared to the standard synthetic control and pre-testing methods

as the number of covariates increases. This indicates that our approach effectively handles the challenges posed by high-dimensional covariate spaces.

To evaluate selection performance of our method, we report True Positive Rate (TPR) and False Positive Rate (FPR). TPR is defined as the proportion of correctly identified valid covariates among all valid covariates, while FPR is the proportion of selected invalid covariates as valid covariates among all invalid covariates. Table 5 summarizes the selection performance across different scenarios. Our method achieves high TPR and low FPR, indicating its effectiveness in accurately distinguising between valid and invalid covariates. Note that as the number of control unit increases, the TPR of our method increases while that of pre-testing method drastically decreases. This again highlights the robustness of our method in poor degree of freedom settings.

4 Conclusion

We propose a new two-step synthetic control method that selects valid covariates using the group square-root LASSO. Using time-variant covariates allows us to penalize synthetic weight directly. Employing the group square-root LASSO enables us to select valid covariates whose weights are the same as outcome variable's weight. We establish the model selection consistency of our method and show that the Post-LASSO estimator achieves the oracle convergence rate under perfect selection. Simulation studies demonstrate the superior performance of our method compared to existing approaches.

While our method shows promising results, several shortcomings remain. First, although our first-stage estimator employs the group square-root LASSO which does not require knowledge of the noise level, the second-stage Post-LASSO estimator still depends on the Gaussian noise assumption. Future research could explore the methods to relax this assumption. Second, our theoretical results focus on the scenario where the number of covariates is less than the number of pre-treatment time periods. Extending the theoretical framework to

accommodate high-dimensional settings where the number of covariates exceeds the number of time periods would be a valuable direction for future work.

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. (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. Papers 2203.06279, arXiv.org.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405.

Andrews, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3):683–734.

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.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-

dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

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*, 84(2):351–381.

Botosaru, I. and Ferman, B. (2019). On the role of covariates in the synthetic control method. *The Econometrics Journal*, 22(2):117–130.

Bunea, F., Lederer, J., and She, Y. (2014). The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory*, 60:1313–1325.

Cavaliere, G., Nielsen, H. B., Pedersen, R. S., and Rahbek, A. (2022). Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models. *Journal of Econometrics*, 227(1):241–263. Annals Issue: Time Series Analysis of Higher Moments and Distributions of Financial Data.

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.

Ellison, M., Lee, S. S., and O'Rourke, K. H. (2024). The ends of 27 big depressions. *American Economic Review*, 114(1):134–68.

Ferman, B. and Pinto, C. (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221.

Ferman, B., Pinto, C., and Possebom, V. (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management*, 39(2):510–532.

Funke, M., Schularick, M., and Trebesch, C. (2023). Populist leaders and the economy. *American Economic Review*, 113(12):3249–88.

Imbens, G. and Rubin, D. (2015). *Causal inference for statistics, social, and biomedical*

sciences: An introduction. Cambridge University Press, New York.

Kellogg, M., Mogstad, M., Pouliot, G. A., and Torgovitsky, A. (2021). Combining matching and synthetic control to tradeoff biases from extrapolation and interpolation. *Journal of the American Statistical Association*, 116(536):1804–1816.

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.

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

Vives-i Bastida, J. (2022). Predictor selection for synthetic controls.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Tables

Table 1: Results for the first scenario: MSE of Synthetic Weight Estimation

Method	$J = 11$	$J = 26$	$J = 41$
Infeasible	0.203	0.239	0.246
Park & Yoon	0.253	0.278	0.287
SCM	0.364	0.347	0.343
Two-Step	0.301	0.345	0.346
w/o Cov	0.328	0.349	0.347

Note: The table reports the MSE of synthetic weight estimation across different methods and degree of freedom scenarios. The "Infeasible" method represents the oracle estimator with knowledge of valid covariates. "Park & Yoon" refers to our proposed two-stage method using group square-root LASSO. "Pre-testing" method represents the two-stage hypothesis testing approach as discussed in section 3. "SCM" denotes the standard synthetic control method, while "w/o Cov" indicates the synthetic control method without covariates.

Table 2: Results for the first scenario: Bias of Synthetic Weight Estimation

Method	$J = 11$	$J = 26$	$J = 41$
Infeasible	0.017	0.012	0.015
w/o Cov	0.021	0.018	0.022
Two-Step	0.029	0.018	0.022
Park & Yoon	0.041	0.040	0.061
SCM	0.330	0.297	0.287

Note: The table reports the RMSE of synthetic weight estimation across different methods and degree of freedom scenarios. The "Infeasible" method represents the oracle estimator with knowledge of valid covariates. "Park & Yoon" refers to our proposed two-stage method using group square-root LASSO. "Pre-testing" method represents the two-stage hypothesis testing approach as discussed in section 3. "SCM" denotes the standard synthetic control method, while "w/o Cov" indicates the synthetic control method without covariates.

Table 3: Results for the second scenario: MSE of Synthetic Weight Estimation

Method	$q = 2$	$q = 6$	$q = 10$
Infeasible	0.261	0.203	0.166
Park & Yoon	0.284	0.253	0.233
Two-Step	0.321	0.301	0.277
w/o Cov	0.325	0.328	0.327
SCM	0.321	0.364	0.362

Note: The table reports the MSE of synthetic weight estimation across different methods and number of covariates scenarios. The "Infeasible" method represents the oracle estimator with knowledge of valid covariates. "Park & Yoon" refers to our proposed two-stage method using group square-root LASSO. "Pre-testing" method represents the two-stage hypothesis testing approach as discussed in section 3. "SCM" denotes the standard synthetic control method, while "w/o Cov" indicates the synthetic control method without covariates.

Table 4: Results for the second scenario: Bias of Synthetic Weight Estimation

Method	$q = 2$	$q = 6$	$q = 10$
Infeasible	0.017	0.017	0.012
Two-Step	0.015	0.029	0.017
w/o Cov	0.014	0.021	0.020
Park & Yoon	0.014	0.041	0.037
SCM	0.230	0.330	0.341

Note: The table reports the RMSE of synthetic weight estimation across different methods and number of covariates scenarios. The "Infeasible" method represents the oracle estimator with knowledge of valid covariates. "Park & Yoon" refers to our proposed two-stage method using group square-root LASSO. "Pre-testing" method represents the two-stage hypothesis testing approach as discussed in section 3. "SCM" denotes the standard synthetic control method, while "w/o Cov" indicates the synthetic control method without covariates.

Table 5: True Positive Rate and False Positive Rate Comparison

Case	Metric	Park & Yoon	Pre-testing
$J = 11$	TPR	0.4613	0.3520
	FPR	0.0187	0.0253
$J = 26$	TPR	0.5480	0.0827
	FPR	0.0987	0.0040
$J = 41$	TPR	0.6293	0.0027
	FPR	0.2120	0.0000
$q = 2$	TPR	0.6520	0.3640
	FPR	0.0440	0.0280
$q = 6$	TPR	0.4613	0.3520
	FPR	0.0187	0.0253
$q = 10$	TPR	0.3832	0.3536
	FPR	0.0192	0.0168

Note: The table presents the True Positive Rate (TPR) and False Positive Rate (FPR) for our proposed method ("Park & Yoon") and the pre-testing method ("Pre-testing") across different simulation scenarios. TPR measures the proportion of correctly identified valid covariates, while FPR indicates the proportion of incorrectly identified invalid covariates as valid ones.

Appendix

A Proof of Theorem 1

Proof. For easy discussion, we decompose the Gram matrix $Z_0^\top Z_0$ into submatrices with $\Sigma_{SS} = \frac{Z_0^{(S)\top} Z_0^{(S)}}{n}$, $\Sigma_{SC} = \frac{Z_0^{(S)\top} Z_0^{(S^C)}}{n}$, $\Sigma_{CS} = \frac{Z_0^{(S^C)\top} Z_0^{(S)}}{n}$, and $\Sigma_{CC} = \frac{Z_0^{(S^C)\top} Z_0^{(S^C)}}{n}$. Following Bunea et al. (2014), we introduce a notation for further discussion:

$$\zeta_{\|\cdot\|_\infty} = \max_{\nu:\|\nu^k\|_2 \le \tilde{\omega}^k \sqrt{p^k} 1 \le g \le q} \max \frac{\|(\tilde{\Sigma}_{21}\nu)^g\|_\infty}{\tilde{\omega}^g \sqrt{p^g}}, \tag{23}$$

where $\tilde{\Sigma}_{21} = (0, \Sigma_{21})^\top$. By definition, for orthogonal design matrix, we have $\zeta_{\|\cdot\|_\infty} = 1$. define the event $\mathcal{B} = \{\|\varepsilon\|_2/\sqrt{n} \le \sqrt{1+\upsilon}\}$ where $\upsilon = \sqrt{\frac{4\log(1/\alpha)}{n}} + \frac{4\log(1/\alpha)}{n}$ for $\alpha \in (0,1)$. Then, we have $P(\mathcal{B}) \ge 1-\alpha$. On the set $\mathcal{A}_1 \cap \mathcal{B}$, we have $\widehat{\omega}^{S^C} = 0$.

Bunea et al. (2014) show that $Z_1 - Z_0\widehat{\omega} \neq 0$ on $\mathcal{A} \cup \mathcal{B}$, and we can consequently apply the Karash-Kuhn-Tucker (KKT) conditions for $\mathcal{C} = \mathcal{A} \cap \mathcal{B}$. Moreover, by definition, $\mathcal{A}_1 \cup \mathcal{B} \subseteq \mathcal{A} \cup \mathcal{B}$, so the results also hold on the smaller set. Thus, there exists a vector $\vartheta \in \mathbb{R}^q$ such that $\|\vartheta^k\|_2 \le \tilde{\omega}^k \sqrt{p^k}$ for all $k = 1, \ldots, q$ and additionally $\vartheta^k = \lambda \tilde{\omega}^k \sqrt{p^k} \frac{\widehat{\omega}^k}{\|\widehat{\omega}^k\|_2}$ when $\widehat{\omega}^k \neq 0$, satisfying the KKT conditions:

$$\frac{Z_0^\top (Z_1 - Z_0\widehat{\omega})}{\|Z_1 - Z_0\widehat{\omega}\|_2} = \frac{1}{n}\vartheta.$$

Denote $\widehat{\Omega} = \|Z_1 - Z_0\widehat{\omega}\|_2$ and $\widehat{\Delta} = \widehat{\omega} - \omega^*$. Then, we have

$$Z_0^\top (\sigma\varepsilon - Z_0\widehat{\Delta}) = \frac{\widehat{\Omega}\vartheta}{\sqrt{n}}. \tag{24}$$

Then, we can rewrite the KKT conditions as:

$$Z_0^{(S)\top}(\sigma\varepsilon - Z_0^{(S)}\widehat{\Delta}^{(S)} - Z_0^{(S^C)}\widehat{\Delta}^{(S^C)}) = \frac{\widehat{\Omega}\vartheta^{(S)}}{\sqrt{n}}, \tag{25}$$

$$Z_0^{(S^C)\top}(\sigma\varepsilon - Z_0^{(S)}\widehat{\Delta}^{(S)} - Z_0^{(S^C)}\widehat{\Delta}^{(S^C)}) = \frac{\widehat{\Omega}\vartheta^{(S^C)}}{\sqrt{n}}. \tag{26}$$

Rearranging the first equation gives:

$$\Sigma_{SS}\widehat{\Delta}^{(S)} + \Sigma_{SC}\widehat{\Delta}^{(S^C)} = \frac{\sigma Z_0^{(S)\top}\varepsilon}{n} - \frac{\widehat{\Omega}\vartheta^{(S)}}{n^{3/2}} \tag{27}$$

$$\Sigma_{CS}\widehat{\Delta}^{(S)} + \Sigma_{CC}\widehat{\Delta}^{(S^C)} = \frac{\sigma Z_0^{(S^C)\top}\varepsilon}{n} - \frac{\widehat{\Omega}\vartheta^{(S^C)}}{n^{3/2}}. \tag{28}$$

Multiplying both sides of Equation (27) by $\widehat{\Delta}^{(S^C)\top}\Sigma_{CS}\Sigma_{SS}^{-1}$ and multiplying both sides of Equation (28) by $\widehat{\Delta}^{(S^C)\top}$, we obtain:

$$\widehat{\Delta}^{(S^C)\top}\Sigma_{CS}\widehat{\Delta}^{(S)} + \widehat{\Delta}^{(S^C)\top}\Sigma_{CS}\Sigma_{SS}^{-1}\Sigma_{SC}\widehat{\Delta}^{(S^C)} = \widehat{\Delta}^{(S^C)\top}\Sigma_{CS}\Sigma_{SS}^{-1}\left(\frac{\sigma Z_0^{(S)\top}\varepsilon}{n} - \frac{\widehat{\Omega}\vartheta^{(S)}}{n^{3/2}}\right), \tag{29}$$

$$\widehat{\Delta}^{(S^C)\top}\Sigma_{CS}\widehat{\Delta}^{(S)} + \widehat{\Delta}^{(S^C)\top}\Sigma_{CC}\widehat{\Delta}^{(S^C)} = \widehat{\Delta}^{(S^C)\top}\left(\frac{\sigma Z_0^{(S^C)\top}\varepsilon}{n} - \frac{\widehat{\Omega}\vartheta^{(S^C)}}{n^{3/2}}\right). \tag{30}$$

Subtract the Equation (30) from Equation (29):

$$\begin{aligned}
&\widehat{\Delta}^{(S^C)\top}\left(\Sigma_{CC} - \Sigma_{CS}\Sigma_{SS}^{-1}\Sigma_{SC}\right)\widehat{\Delta}^{(S^C)} \\
&\quad= \widehat{\Delta}^{(S^C)\top}\left(\frac{\sigma Z_0^{(S^C)\top}\varepsilon}{n} - \frac{\widehat{\Omega}\vartheta^{(S^C)}}{n^{3/2}}\right) - \widehat{\Delta}^{(S^C)\top}\Sigma_{CS}\Sigma_{SS}^{-1}\left(\frac{\sigma Z_0^{(S)\top}\varepsilon}{n} - \frac{\widehat{\Omega}\vartheta^{(S)}}{n^{3/2}}\right) \qquad (31) \\
&\quad= \underbrace{\widehat{\Delta}^{(S^C)\top}\left(\frac{\sigma Z_0^{(S^C)\top}(I - P_S)\varepsilon}{n}\right)}_{(A)} - \underbrace{\widehat{\Delta}^{(S^C)\top}\left(\frac{\widehat{\Omega}\vartheta^{(S^C)}}{n^{3/2}} - \Sigma_{CS}\Sigma_{SS}^{-1}\frac{\widehat{\Omega}\vartheta^{(S)}}{n^{3/2}}\right)}_{(B)}, \qquad (32)
\end{aligned}$$

where P_S is a projection matrix. We aim to show that the right-hand side of Equation (31) is negative to show that $\widehat{\Delta}^{S^C} = 0$. Note that $\Sigma_{CC} - \Sigma_{CS}\Sigma_{SS}^{-1}\Sigma_{SC}$ is positive definite. Since P_S is a projection matrix, $\|Z_0^{S^C}(I - P_S)\varepsilon\| \leq \|Z_0^{S^C}\varepsilon\|$. If there exists $M \geq 0$ such that $\|Z_0^{S^C}\varepsilon\| \leq M$, the term (A) is bounded by $M\sum_{k\notin S}\|\widehat{\Delta}^k\|$. For the term (B), we rewrite it

as:

$$\widehat{\Omega}\left(\sum_{k\in S^C}(\widehat{\Delta}^k)^\top\vartheta^k-\sum_{k\in S^C}(\widehat{\Delta}^k)^\top\left[\Sigma_{CS}\Sigma_{SS}^{-1}\vartheta^S\right]^k\right). \tag{33}$$

First, by Lemma 4.3. in Bunea et al. 2014, $\widehat{\Omega}$ is bounded at some constant times Ω. Next, for the second term of Equation (10), we have:

$$\sum_{k\in S^C}(\widehat{\Delta}^k)^\top\left[\Sigma_{CS}\Sigma_{SS}^{-1}\vartheta^S\right]^k\leq\sum_{k\in S^C}\|\widehat{\Delta}^k\|_2\cdot\left\|\left[\Sigma_{CS}\Sigma_{SS}^{-1}\vartheta^S\right]^k\right\|_2 \tag{34}$$

$$\leq\sum_{k\in S^C}\|\widehat{\Delta}^k\|_2\lambda(1-\eta), \tag{35}$$

where $\eta\in(0,1)$. Inequality (12) follows from the Assumption 1 (*Group Irrepresentable Condition*). Thus, the term (B) can be bounded from below as:

$$\sum_{k\in S^C}\|\widehat{\Delta}^k\|_2\lambda-\widehat{\Delta}^{S^C\top}\Sigma_{CS}\Sigma_{SS}^{-1}\vartheta^S\geq\sum_{k\in S^C}\|\widehat{\Delta}^k\|_2\lambda-\sum_{k\in S^C}\|\widehat{\Delta}^k\|_2\lambda(1-\eta) \tag{36}$$

$$=\sum_{k\in S^C}\widehat{\Delta}^k\|_2\lambda\eta. \tag{37}$$

Since $\Sigma_{CC}-\Sigma_{CS}\Sigma_{SS}^{-1}\Sigma_{SC}$ is positive definite while the right-hand side of Equation (8) is negative, we conclude that $\widehat{\Delta}^C=0$. Next, we prove part 2 of the theorem. We start from Equation (24) and using $\widehat{\Delta}^{(S^C)}=0$, we have

$$-n^2\Sigma_{SS}\widehat{\Delta}^{(S)}=\sqrt{n}\widehat{\Omega}\vartheta^{(S)}-n\sigma(Z_0^\top\varepsilon)^{(S)}.$$

It implies that

$$-n^2\widehat{\Delta}^{(S)}=\sqrt{n}\widehat{\Omega}\Sigma_{SS}^{-1}\left(\vartheta^{(S)}-\frac{\sqrt{n}\sigma}{\lambda\widehat{\Omega}}(Z_0^\top\varepsilon)^{(S)}\right).$$

Using $\lambda/\tilde{\eta} \leq \widehat{V}$ where $\widehat{V} = \max_{1\leq k\leq q}\left\{\frac{\sigma\sqrt{n}\|(Z_0^\top\varepsilon)^k\|_2}{\tilde{\omega}^k\sqrt{p^k}\widehat{\Omega}}\right\}$ and bounding the norms as above, we have

$$\begin{aligned}
\|\widehat{\Delta}^k\|_\infty &\leq \max_{\nu:\|\nu^g\|_2\leq\widehat{\omega}^g\sqrt{p^g}} \frac{\left(1+\frac{1}{\tilde{\eta}}\right)}{n}\frac{\widehat{\Omega}}{\sqrt{n}}\|(\tilde{\Sigma}_{SS}^{-1}\nu)^k\|_\infty \\
&\leq \frac{\left(1+\frac{1}{\tilde{\eta}}\right)\tilde{\omega}^k\sqrt{p^k}\lambda}{n}\frac{\widehat{\Omega}}{\sqrt{n}}\zeta_{\|\cdot\|_\infty} \\
&\leq \frac{\left(1+\frac{1}{\tilde{\eta}}\right)\tilde{\omega}^k\sqrt{p^k}\lambda}{n}\frac{\|\sigma\varepsilon\|_2}{\sqrt{n}}\left(1+\frac{\lambda\sqrt{s^*}u}{n\kappa}\right)\zeta_{\|\cdot\|_\infty} \\
&\leq D\frac{\tilde{\omega}^k\sqrt{p^k}\lambda}{n}, \quad \text{for all } 1\leq k\leq q.
\end{aligned}$$

For the third inequality, we have applied Lemma 4.3 in Bunea et al. (2014) to bound $\widehat{\Omega}$. The last claim follows directly from the first two parts of the theorem and Assumption 2. This completes the proof. □

B Proof of Theorem 2

Proof. Since our proposed method assumes same ones as the Post-LASSO procedure in Belloni and Chernozhukov (2013), Theorem 2 can be proved by adapting their proof with minor modifications. Belloni and Chernozhukov (2013) assumes that their penalty parameter λ should dominate the noise, so that they propose to set $\lambda = c'\widehat{\sigma}\Lambda(1-\alpha|X)$ where $\Lambda(1-\alpha|X)$ is the $(1-\alpha)$ quantile of $n\|2\mathbb{E}[X\varepsilon]/\sigma\|_\infty$ for $c' > c > 1$ and $\lambda \geq cn\|2\mathbb{E}[X\varepsilon]\|_\infty$. It is equivalent to assume to discuss our theoretical properties under the set $\mathcal{A}_1$ since it assumes that penalty parameter λ should be chosen to dominate the noise level. Moreover, we have strong theoeretical advantage in terms that our proposed method is pivotal. For completeness, however, we provide other details here.

Let $\widehat{\mathcal{S}}^C$ be the set of selected valid covariates in the first-stage, and define $\mathcal{W}$ as the support of second-stage estimator given $\widehat{\mathcal{S}}^C$. The second-stage Post-LASSO estimator $\widehat{w}^{\text{Post}}$ is defined as the minimizer of Equation (18) supported on $\mathcal{W}$. Note that the true weight w^* might not

be supported on $\mathcal{W}^C$ if the first-stage selection is imperfect.

Let $\tilde{w}$ be the approximation of w^* on the selected support $\mathcal{W}^C$, defined as:

$$\tilde{w} \in \underset{w:supp(w)\subseteq\widehat{\mathcal{W}}}{\arg\min} \|Z_1^{\text{Post}} - Z_0^{\text{Post}} w\|_2^2.$$

By definition, we have

$$\|Z_1^{\text{Post}} - Z_0^{\text{Post}} \widehat{w}^{\text{Post}}\|_2^2 \leq \|Z_1^{\text{Post}} - Z_0^{\text{Post}} \tilde{w}\|_2^2. \tag{38}$$

Substituting $Z_1^{\text{Post}} = Z_0^{\text{Post}} w^* + \varepsilon^{\text{Post}}$ into inequality (38), we have

$$\|Z_0^{\text{Post}}(w^* - \widehat{w}^{\text{Post}}) + \varepsilon^{\text{Post}}\|_2^2 \leq \|Z_0^{\text{Post}}(w^* - \tilde{w}) + \varepsilon^{\text{Post}}\|_2^2.$$

Let $r = Z_0^{\text{Post}}(w^* - \tilde{w})$ and $\tilde{\Delta} = \widehat{w}^{\text{Post}} - \tilde{w}$. Note that $\tilde{\Delta}$ is supported on $\mathcal{W}$. Rearranging the above inequality, we obtain:

$$\|r - Z_0^{\text{Post}} \tilde{\Delta} + \varepsilon^{\text{Post}}\|_2^2 \leq \|r + \varepsilon^{\text{Post}}\|_2^2.$$

Expanding the left-hand side, we have

$$\|r + \varepsilon^{\text{Post}}\|_2^2 - 2(r + \varepsilon^{\text{Post}})^\top Z_0^{\text{Post}} \tilde{\Delta} + \|Z_0^{\text{Post}} \tilde{\Delta}\|_2^2 \leq \|r + \varepsilon^{\text{Post}}\|_2^2.$$

It follows that

$$\|Z_0^{\text{Post}} \tilde{\Delta}\|_2^2 \leq 2(r + \varepsilon^{\text{Post}})^\top Z_0^{\text{Post}} \tilde{\Delta}.$$

Note that by the definition of $\tilde{w}$, we have $r^\top Z_0^{\text{Post}} \tilde{\Delta} = 0$. Therefore, we obtain

$$\|Z_0^{\text{Post}} \tilde{\Delta}\|_2^2 \leq 2\varepsilon^{\text{Post}\top} Z_0^{\text{Post}} \tilde{\Delta}. \tag{39}$$

Now we control the empirical process term on the right-hand side. Let $\mathcal{P}_{Z_0^{\text{Post}}}$ denote the projection matrix onto the column space of Z_0^{Post} corresponding to the selected support $\widehat{\mathcal{S}}^C$, as Z_0^{Post} depends on the first-stage model selection $\widehat{\mathcal{S}}^C$. Since $\tilde{\Delta}$ lies in this subspace, we have

$$\varepsilon^\top Z_0^{\text{Post}} \tilde{\Delta} = \|\mathcal{P}_{Z_0^{\text{Post}}} \varepsilon\|_2 \|Z_0^{\text{Post}} \tilde{\Delta}\|_2,$$

and we get

$$\|Z_0^{\text{Post}}(\widehat{w}^{\text{Post}} - \tilde{w})\|_2 \leq 2\|\mathcal{P}_{Z_0^{\text{Post}}} \varepsilon^{\text{Post}}\|_2.$$

Using triangular inequality, we have

$$\begin{aligned}\|Z_0^{\text{Post}}(\widehat{w}^{\text{Post}} - w^*)\|_2 &\leq \|Z_0^{\text{Post}}(\widehat{w}^{\text{Post}} - \tilde{w})\|_2 + \|Z_0^{\text{Post}}(\tilde{w} - w^*)\|_2 \\ &\leq 2\|\mathcal{P}_{Z_0^{\text{Post}}} \varepsilon^{\text{Post}}\|_2 + \|r\|_2.\end{aligned}$$

By the maximal inequality for projections onto sparse subspaces, we have with high probability

$$\|\mathcal{P}_{Z_0^{\text{Post}}} \varepsilon^{\text{Post}}\|_2 \leq \sup_{\mathcal{I} \in \mathcal{V}} \|\mathcal{P}_{\mathcal{I}} \varepsilon_{\mathcal{I}}^{\text{Post}}\|_2 \leq \sigma^{\text{Post}}(\sqrt{J} + \sqrt{C_{\tilde{K}} \dot{s} \log q})$$

where $\mathcal{V} = \{\mathcal{I} \subseteq \{1, \ldots, q\} : |\mathcal{I}^C| \leq C_{\tilde{K}} \dot{s}\}$ for some constant $C_{\tilde{K}} > 0$ by Theorem 3 in Belloni and Chernozhukov (2013). Here, we bound the dimension of the selected subspace to potential valid covariates sets $\mathcal{V}$. Thus, we obtain

$$\frac{1}{\sqrt{n^{Post}}} \|Z_0^{\text{Post}}(\widehat{w}^{\text{Post}} - w^*)\|_2 \leq 2\sigma^{\text{Post}} \left(\frac{\sqrt{J} + \sqrt{C_{\tilde{K}} \dot{s} \log q}}{\sqrt{n^{Post}}} \right) + \frac{\|r\|_2}{\sqrt{n^{Post}}}.$$

This completes the proof and it can be also applied to the case where the weight vector is constrained to be non-negative and sum to one. □

C Proof of Corollary 1

Proof. The proof consists of establishing the selection consistency and then deriving the oracle property as a direct consequence.

1. **Selection Consistency:** From Theorem 1, we know that under the specified assumptions, the first-stage estimator $\widehat{\omega}$ achieves model selection consistency. Specifically, with probability at least $1-\alpha$, if the true non-zero biases are sufficiently large (as per Assumption 2), the estimated support $\widehat{\mathcal{S}}^C$ will coincide with the true support $\mathcal{S}^C$ as $n \to \infty$. Since α can be chosen to decay to 0 as $n \to \infty$ provided that the tuning parameter λ is selected appropriately (see Section 2.4), we have

$$\lim_{n\to\infty} P(\widehat{\mathcal{S}}^C = \mathcal{S}^C) = 1.$$

This implies that the set of covariates identified as *valid*, those in $\widehat{\mathcal{S}}^C$, is exactly the set of truly valid covariates $\mathcal{S}^C$ asymptotically.

2. **Oracle Property:** The second-stage estimator $\widehat{w}^{\text{Post}}$ is obtained by solving the least squares problem (18) using the design matrices Z_0^{Post} that includes only the covariates in $\widehat{\mathcal{S}}^C$. The oracle estimator $\widehat{w}^{\text{Oracle}}$ is defined as the same optimization problem but using the design matrices Z_1^{Oracle} and Z_0^{Oracle} that include only the truly valid covariates in $\mathcal{S}^C$.

Consider the event $\mathcal{E} = \{\widehat{\mathcal{S}} = \mathcal{S}\}$. On this event, the selected set of covariates is identical to the true set of valid covariates. Consequently, the design matrices are identical: $Z_1^{\text{Post}} = Z_1^{\text{Oracle}}$ and $Z_0^{\text{Post}} = Z_0^{\text{Oracle}}$. Therefore, on the event $\mathcal{E}$, the Post-LASSO estimator $\widehat{w}^{\text{Post}}$ coincides with the oracle estimator $\widehat{w}^{\text{Oracle}}$.

Since we have already established that $P(\mathcal{E}) \to 1$ as $n \to \infty$, it follows that

$$P(\widehat{w}^{\text{Post}} = \widehat{w}^{\text{Oracle}}) \to 1 \quad \text{as } n \to \infty.$$

3. **Convergence Rate:** On the event $\mathcal{E}$, we can apply Theorem 2, and the cardinality of the selected covariates $\widehat{q}$ equals the true number of valid covariates $q - \dot{s}$. We have:

$$\frac{1}{\sqrt{n}}\|Z_0^{\text{Oracle}}(\widehat{w}^{\text{Post}} - w^*)\|_2 \precsim \sigma\sqrt{\frac{\dot{s}}{n}}.$$

where $a \precsim b$ denotes $a = O_p(b)$. This completes the proof of the corollary. □