SAME ROOT DIFFERENT LEAVES: TIME SERIES AND CROSS-SECTIONAL METHODS IN PANEL DATA

DENNIS SHEN

Simons Institute for the Theory of Computing, University of California, Berkeley

PENG DING

Department of Statistics, University of California, Berkeley

JASJEET SEKHON

Departments of Statistics & Data Science and Political Science, Yale University

BIN YU

Departments of Statistics and EECS, University of California, Berkeley

A central goal in social science is to evaluate the causal effect of a policy. One dominant approach is through panel data analysis in which the behaviors of multiple units are observed over time. The information across time and space motivates two general approaches: (i) horizontal regression (i.e., unconfoundedness), which exploits time series patterns, and (ii) vertical regression (e.g., synthetic controls), which exploits cross-sectional patterns. Conventional wisdom states that the two approaches are fundamentally different. We establish this position to be partly false for estimation but generally true for inference. In particular, we prove that both approaches yield identical point estimates under several standard settings. For the same point estimate, however, each approach quantifies uncertainty with respect to a distinct estimand. The confidence interval developed for one estimand may have incorrect coverage for another. This emphasizes that the source of randomness that researchers assume has direct implications for the accuracy of inference.

KEYWORDS: horizontal regression, vertical regression, unconfoundedness, synthetic controls, causal inference, minimum norm estimators.

1. INTRODUCTION

In a seminal paper, Abadie and Gardeazabal (2003) set out to investigate the economic impact of terrorism in Basque Country. Prior to the outset of terrorist activity in the early 1970's, Basque Country was considered to be one of the wealthiest regions in Spain. After thirty years of turmoil, however, its economic activity dropped substantially relative to its neighboring regions. Although intuition affirms that Basque Country's economic downturn can be attributed, at least partially, to its political and civil unrest, it is difficult to quantitatively isolate the economic costs of conflict. In response to this challenge, Abadie and Gardeazabal (2003) introduced the synthetic

Dennis Shen: dshen24@berkeley.edu

Peng Ding: pengdingpku@berkeley.edu

Jasjeet Sekhon: jasjeet.sekhon@yale.edu

Bin Yu: binyu@berkeley.edu
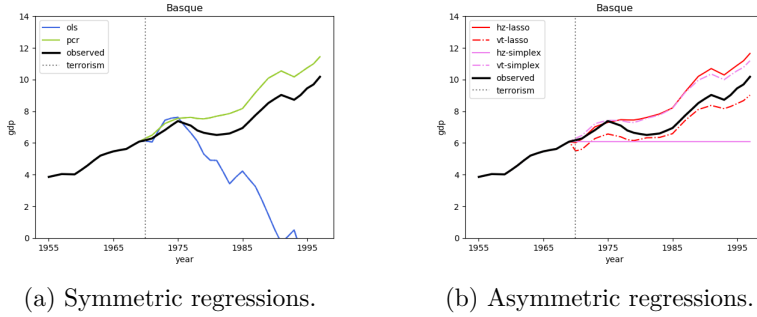
(a) Symmetric regressions.

(b) Asymmetric regressions.

FIGURE 1.—1a: Estimates of OLS with minimum ℓ_2-norm and PCR. 1b: Estimates of lasso and simplex regression. HZ and VT estimates correspond to colored solid and dashed-dotted lines, respectively. The outset of terrorism is the vertical line and Basque Country's observed GDP is in solid black.

controls framework. At its core, synthetic controls constructs a synthetic Basque Country from a weighted composition of control regions that are largely unaffected by the instability to estimate Basque Country's economic evolution in the absence of terrorism. This novel concept has inspired an entire subliterature within econometrics that is "arguably the most important innovation in the policy evaluation literature in the last 15 years" (Athey and Imbens, 2017).

Researchers have historically tackled problems of this flavor using repeated observations of units across time, i.e., panel data, where a subset of units are exposed to a treatment during some time periods while the other units are unaffected. In the study above, the per capita gross domestic product (GDP) of 17 Spanish regions are measured from 1955–1998. Basque Country is the sole treated unit and the remaining regions are the control units; the pre- and post-treatment periods are defined as the time horizons before and after the first wave of terrorist activity, respectively.

Synthetic controls has become a cornerstone for panel studies in recent years and across numerous fields. Beforehand, the unconfoundedness approach (Rosenbaum and Rubin, 1983, Imbens and Wooldridge, 2009) served as a common workhorse. Whereas synthetic controls posits a relation between treated and control units that is stable across time, unconfoundedness posits a relation between treated and pretreatment periods that is stable across units. Accordingly, synthetic controls exploits cross-sectional correlation patterns while unconfoundedness exploits time series correlation patterns. Considering the panel data format, unconfoundedness and synthetic controls based methods are commonly referred to as horizontal (HZ) and vertical (VT) regressions, respectively. Given their conceptual and computational distinctions, the two approaches are considered to be fundamentally different (Athey et al., 2021).

Yet, contrary to conventional wisdom, it turns out that HZ and VT regressions can yield identical point estimates. As Figure 1a shows, when the regression models are learned via ordinary least squares (OLS) or principal component regression (PCR), then the two approaches produce the same economic evolution for Basque Country in the absence of terrorism. Figure 1b, by contrast, shows that when the regression models are learned via lasso or lie within the simplex—as proposed by Abadie and Gardeazabal (2003) for VT regression—then the two approaches output contrasting economic trajectories. Curiously, Figure 2 indicates that even when the two regressions arrive at the same point estimate, the confidence intervals can be markedly different under different sources of randomness. The juxtaposition of these figures beg two questions:

(a) OLS under HZ model.

(b) OLS under VT model.

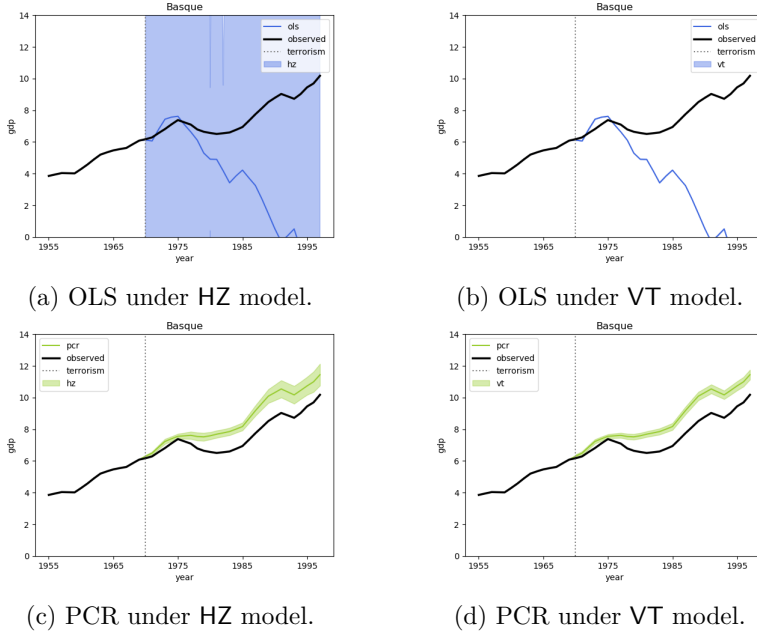(c) PCR under HZ model.

(d) PCR under VT model.

FIGURE 2.—Confidence intervals for OLS (top) and PCR (bottom) constructed from HZ-based (left) and VT-based (right) generative models, which are precisely defined in Section 4.1.1.

Q1: "When are HZ and VT point estimates identical?"
Q2: "When the point estimates are identical, how does the source of randomness impact inference?"

Contribution. This article tackles Q1–Q2 from first principles. In this endeavor, we begin by classifying several widely studied regression formulations into (i) a symmetric class that yields identical point estimates and (ii) an asymmetric class that yields contrasting point estimates. Within the symmetric class, we study properties of the estimator with randomness stemming from (i) time series patterns, (ii) cross-sectional patterns, and (iii) both patterns. We conduct our analysis from a (i) model-based perspective, which attributes randomness to the potential outcomes, and a (ii) design-based perspective, which attributes randomness to the treatment assignment mechanism. In both frameworks, we find that the source of randomness has large implications for the estimand and inference. Under the model-based framework, we construct confidence intervals for each source of randomness. Through data-inspired simulations and empirical applications, we demonstrate that the confidence interval developed for one estimand often has incorrect coverage for another estimand. Taken together, our results emphasize that the source of randomness that researchers assume has direct implications for the accuracy of the inference that can be conducted.

Organization. Section 2 overviews the panel data framework. Sections 3–4 provide one set of answers for Q1–Q2. Section 5 illustrates concepts developed in this article. Section 6 summarizes our findings. Details of simulations and empirical applications, select discussions, and mathematical proofs are relegated to the appendix.

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} & \cdots & Y_{1T_0} & Y_{1T} \\ \vdots & \ddots & \vdots & \vdots \\ Y_{N_01} & \cdots & Y_{N_0T_0} & Y_{N_0T} \\ \hline Y_{N1} & \cdots & Y_{NT_0} & ? \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y}_0 & \vline & \boldsymbol{y}_T \\ \hline \boldsymbol{y}'_N & \vline & ? \end{pmatrix} \begin{matrix} \text{control units} \\ \\ \text{treated unit} \end{matrix}$$

$$\underbrace{\phantom{Y_{11} \cdots Y_{1T_0}}}_{\text{pretreatment period}} \underbrace{\phantom{Y_{1T}}}_{\text{treated period}} \qquad \underbrace{\phantom{\boldsymbol{Y}_0}}_{\text{pretreatment period}} \underbrace{\phantom{\boldsymbol{y}_T}}_{\text{treated period}}$$

FIGURE 3.—Panel data format with rows and columns indexed by units and time, respectively.

Notation. Let $\boldsymbol{I}$ be the identity matrix. Let $\boldsymbol{1}$ and $\boldsymbol{0}$ be the vectors of ones and zeros, respectively. The curled inequality denotes $\succeq$ the generalized inequality, i.e., componentwise inequality between vectors and matrix inequality between symmetric matrices. Let $\circ$ denote the componentwise product. For vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, let $\langle a, b \rangle = a'b$ denote the inner product. Let $\operatorname{tr}(\boldsymbol{A})$ denote the trace of $\boldsymbol{A}$. We define $0/0 = 0$ when applicable.

2. THE PANEL DATA FRAMEWORK

We anchor on the Basque study to introduce the panel data framework. Panel data contains observations of N units over T time periods. The Basque study, for instance, consists of per capita GDP across $N = 17$ Spanish regions over $T = 43$ years. In each time period t, each unit i is characterized by two potential outcomes, $Y_{it}(0)$ and $Y_{it}(1)$, which correspond to its outcome in the absence and presence of a binary treatment, respectively. The potential outcomes framework posits that each region possesses two possible levels of economic activity each year, one that is immune to terrorism and another that is affected by terrorism. In reality, however, we can only observe one economic state, $Y_{it}(0)$ or $Y_{it}(1)$—this is the fundamental challenge of causal inference.

Let Y_{it} be the observed outcome. Often, we observe all N units without treatment (control) for T_0 time periods, i.e., $Y_{it} = Y_{it}(0)$ for all $i \leq N$ and $t \leq T_0$. For the remaining $T_1 = T - T_0$ time periods, N_1 units receive treatment while the remaining $N_0 = N - N_1$ units remain under control, i.e., if we arbitrarily label the first N_0 units as the control group, then $Y_{it} = Y_{it}(1)$ for all $i > N_0$ and $t > T_0$, and $Y_{it} = Y_{it}(0)$ for all $i \leq N_0$ and $t > T_0$. In our study, Basque Country is the single treated unit, thus $N_1 = 1$ and $N_0 = 16$. The first wave of terrorist activity partitions the time horizon into pre- and post-treatment periods of lengths $T_0 = 15$ and $T_1 = 28$ years, respectively.

For ease of exposition, this article considers a single treated unit and single treated period indexed by the Nth unit and Tth time period, respectively. However, our results hold for any (i, t) pair where $i > N_0$ is a treated unit and $t > T_0$ is a treated period. We organize our observed control data into an $N \times T$ matrix, $\boldsymbol{Y} = [Y_{it}]$, as shown in Figure 3. In our example, $\boldsymbol{y}_N = [Y_{Nt} : t \leq T_0] \in \mathbb{R}^{T_0}$ represents Basque Country's economic evolution prior to the outset of terrorism; $\boldsymbol{Y}_0 = [Y_{it} : i \leq N_0, t \leq T_0] \in \mathbb{R}^{N_0 \times T_0}$ represents the control regions' economic evolution prior to the outset of terrorism; and $\boldsymbol{y}_T = [Y_{iT} : i \leq N_0] \in \mathbb{R}^{N_0}$ represents the control regions' economic evolution after the outset of terrorism. Our object of interest is Basque Country's counterfactual GDP in the absence of terrorism, $Y_{NT}(0)$.

2.1. *Time Series Versus Cross-Sectional Based Regressions*

The information across time and space motivates two natural ways to impute the missing (N, T)th entry. These perspectives are explored in two large and mostly separate bodies of work (Athey et al., 2021).

2.1.1. *Horizontal Regression and Unconfoundedness*

The unconfoundedness literature operates on the concept that "history is a guide to the future". As such, unconfoundedness methods express outcomes in the treated period as a weighted composition of outcomes in the pretreatment periods. This is carried out by regressing the control units' treated period outcomes $\boldsymbol{y}_T$ on its lagged outcomes $\boldsymbol{Y}_0$ and applying the learned regression coefficients to the treated unit's lagged outcomes $\boldsymbol{y}_N$ to predict the missing (N, T)th outcome. Following Athey et al. (2021), we refer to such methods as horizontal (HZ) regression.

2.1.2. *Vertical Regression and Synthetic Controls*

The synthetic controls literature is built on the concept that "similar units behave similarly". Therefore, synthetic controls methods express the treated unit's outcomes as a weighted composition of control units' outcomes. This is carried out by regressing the treated unit's lagged outcomes $\boldsymbol{y}_N$ on the control units' lagged outcomes $\boldsymbol{Y}_0'$ and applying the learned regression coefficients to the control units' treated period outcomes $\boldsymbol{y}_T$ to predict the missing (N, T)th outcome. Following Athey et al. (2021), we refer to such methods as vertical (VT) regression.

2.1.3. *Conventional Wisdom*

The asymmetry between HZ and VT regressions has created the conception that they are fundamentally different approaches (Athey et al., 2021). In fact, the unregularized forms of HZ and VT regressions are cautioned against when $T > N$ and $N > T$, respectively (Abadie et al., 2015, Doudchenko and Imbens, 2016, Li and Bell, 2017, Athey et al., 2021). With regularization, however, Athey et al. (2021) argues the two approaches can be applied to the same setting. In turn, this allows the two approaches to be systematically compared through methods such as cross-validation.

In parallel, the growth rates of the two literatures have also exhibited asymmetry. While the development of the unconfoundedness literature has seemingly plateaued, the synthetic controls literature continues to rapidly expand. Across many domains, synthetic controls based methods are arguably the de facto approach for panel studies.

3. POINT ESTIMATION

> **Q1:** "When are HZ and VT point estimates identical?"

We tackle Q1 by studying the finite-sample estimation properties of HZ and VT regressions. We denote the singular value decomposition of $\boldsymbol{Y}_0$ as $\boldsymbol{Y}_0 = \sum_{\ell=1}^{R} s_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell' = \boldsymbol{U} \boldsymbol{S} \boldsymbol{V}'$, where $\boldsymbol{u}_\ell \in \mathbb{R}^{N_0}$ and $\boldsymbol{v}_\ell \in \mathbb{R}^{T_0}$ are the left and right singular vectors, respectively, $s_\ell \in \mathbb{R}$ are the ordered singular values, and $R = \mathrm{rank}(\boldsymbol{Y}_0) \leq \min\{N_0, T_0\}$. $\boldsymbol{U} \in \mathbb{R}^{N_0 \times R}$ and $\boldsymbol{V} \in \mathbb{R}^{T_0 \times R}$ denote the matrices formed by the left and right singular vectors, respectively, and $\boldsymbol{S} \in \mathbb{R}^{R \times R}$ is the diagonal matrix of singular values. The Moore-Penrose pseudoinverse of $\boldsymbol{Y}_0$ is $\boldsymbol{Y}_0^\dagger = \sum_{\ell=1}^{R} (1/s_\ell) \boldsymbol{v}_\ell \boldsymbol{u}_\ell' = \boldsymbol{V} \boldsymbol{S}^{-1} \boldsymbol{U}'$. Critically, we do not place any assumptions on the relative magnitudes of N and T.

3.1. *Classifying Notable Regression Formulations*

We present several of the most widely studied regression formulations in the HZ and VT literatures. This list is far from exhaustive given the vastness of these literatures.

3.1.1. *Description of Estimation Strategies*

Penalized regression. A large class of penalized regressions are expressed as follows:

(a) HZ regression: for $\lambda_1, \lambda_2 \geq 0$,

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\text{argmin}} \ \|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2 \tag{1}$$

$$\widehat{Y}_{NT}^{\text{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle.$$

(b) VT regression: for $\lambda_1, \lambda_2 \geq 0$,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \ \|\boldsymbol{y}_N - \boldsymbol{Y}_0'\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \tag{2}$$

$$\widehat{Y}_{NT}^{\text{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle.$$

We overview common choices for (λ_1, λ_2) and describe the corresponding strategy.

I: Ordinary least squares (OLS). Arguably, the mother of all regressions is OLS, where $\lambda_1 = \lambda_2 = 0$. OLS is an unconstrained problem with possibly infinitely many solutions. OLS has been analyzed in numerous works on panel studies, including Hsiao et al. (2012), Li and Bell (2017), and Li (2020).

II: Principal component regression (PCR). To formalize PCR, let

$$\boldsymbol{Y}_0^{(k)} = \sum_{\ell=1}^k s_\ell \boldsymbol{u}_\ell \boldsymbol{v}_\ell' \tag{3}$$

denote the rank $k < R$ approximation of $\boldsymbol{Y}_0$ that retains the top k principal components. HZ and VT PCR corresponds to replacing $\boldsymbol{Y}_0$ with $\boldsymbol{Y}_0^{(k)}$ within (1) and (2), respectively, with $\lambda_1 = \lambda_2 = 0$. In words, PCR first finds a k dimensional representation of the covariate matrix via principal component analysis; then, PCR performs OLS with the compressed k dimensional covariates. Within the synthetic controls literature, Amjad et al. (2018, 2019) and Agarwal et al. (2021) utilize PCR.

III: Ridge regression. Consider ridge regression, where $\lambda_1 = 0$ and $\lambda_2 > 0$. When $\boldsymbol{Y}_0$ is rank deficient, the gram matrix, i.e., $\boldsymbol{Y}_0'\boldsymbol{Y}_0$ for HZ regression and $\boldsymbol{Y}_0\boldsymbol{Y}_0'$ for VT regression, is ill-conditioned. This often discourages the usage of OLS. In these settings, ridge provides a remedy by adding a *ridge* on the diagonal of the gram matrix, which increases all eigenvalues by λ_2, thus removing the singularity problem. Ben-Michael et al. (2021) explores the properties of a doubly robust estimator that utilizes HZ ridge regression.

IV: Lasso regression. Consider lasso regression, where $\lambda_1 > 0$ and $\lambda_2 = 0$. Lasso has become a popular tool for estimating sparse linear coefficients in high-dimensional regimes. Because the lasso criterion not strictly convex, there are possibly infinitely many solutions. Thus, for our analysis of lasso only, we make the mild assumption that the entries of $\boldsymbol{Y}_0$ are drawn from a continuous distribution. As established in Tibshirani (2013),

this guarantees the lasso solution to be unique. Several notable works in the synthetic controls literature, e.g., Li and Bell (2017), Carvalho et al. (2018), and Chernozhukov et al. (2021), analyze the lasso.

V: Elastic net regression. Mixing both ℓ_1 and ℓ_2-penalties, i.e., $\lambda_1, \lambda_2 > 0$, is known as elastic net. At a high level, elastic net selects variables similar to the lasso, but deals with correlated variables more gracefully as with ridge. When $\lambda_2 > 0$, the criterion is strictly convex so the solution is unique. Doudchenko and Imbens (2016) propose an elastic net synthetic controls variant.

Constrained regression.
VI: Simplex regression. The next formulation constrains the regression weights to lie within the simplex, i.e., the weights are nonnegative and sum to one:

(a) HZ regression: for $\lambda \geq 0$,

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \ \|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 \ \text{ subject to } \ \boldsymbol{\alpha}'\mathbf{1} = 1, \boldsymbol{\alpha} \succeq \mathbf{0} \qquad (4)$$

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle.$$

(b) VT regression: for $\lambda \geq 0$,

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \ \|\boldsymbol{y}_N - \boldsymbol{Y}_0'\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \ \text{ subject to } \ \boldsymbol{\beta}'\mathbf{1} = 1, \boldsymbol{\beta} \succeq \mathbf{0} \qquad (5)$$

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle.$$

We consider a vanishing ℓ_2 penalty since $\lambda = 0$ (standard formulation) can induce multiple minima (Abadie and L'Hour, 2021). When $\lambda > 0$, the criterion becomes strictly convex and the solution is unique. Simplex regression is the original formulation set forth in the pioneering works of Abadie and Gardeazabal (2003), Abadie et al. (2010, 2015), and its properties continue to be actively studied today. Attractive aspects of simplex regression include interpretability, sparsity, and transparency (Abadie, 2021).

3.1.2. *Classification Results*

To answer Q1, we classify the regression formulations into (i) a symmetric class, where HZ and VT point estimates agree, and (ii) an asymmetric class, where HZ and VT point estimates disagree. We use the shorthand $\mathsf{HZ} = \mathsf{VT}$ if the two approaches produce identical point estimates and $\mathsf{HZ} \neq \mathsf{VT}$ otherwise.

I: Symmetric class. We first state the symmetric formulations.

THEOREM 1: $\mathsf{HZ} = \mathsf{VT}$ *for (i) OLS with* $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ *as the minimum* ℓ_2-*norm solutions:*

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_N, \boldsymbol{Y}_0^{\dagger}\boldsymbol{y}_T \rangle = \sum_{\ell=1}^{R} (1/s_\ell)\langle \boldsymbol{y}_N, \boldsymbol{v}_\ell \rangle \langle \boldsymbol{u}_\ell, \boldsymbol{y}_T \rangle;$$

(ii) PCR with the same choice of $k < R$:

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0^{(k)})^{\dagger}\boldsymbol{y}_T \rangle = \sum_{\ell=1}^{k} (1/s_\ell)\langle \boldsymbol{y}_N, \boldsymbol{v}_\ell \rangle \langle \boldsymbol{u}_\ell, \boldsymbol{y}_T \rangle;$$

(iii) ridge regression with the same choice of $\lambda_2 > 0$:

$$\widehat{Y}_{NT}^{\text{hz}}(0) = \widehat{Y}_{NT}^{\text{vt}}(0) = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0' \boldsymbol{Y}_0 + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{Y}_0' \boldsymbol{y}_T \rangle = \sum_{\ell=1}^{R} \frac{s_\ell}{s_\ell^2 + \lambda_2} \langle \boldsymbol{y}_N, \boldsymbol{v}_\ell \rangle \langle \boldsymbol{u}_\ell, \boldsymbol{y}_T \rangle.$$

Theorem 1 might seem familiar at first glance. As observed in Abadie et al. (2015) and Ben-Michael et al. (2021, Lemma 2), the point estimates associated with HZ OLS and HZ ridge can be written as linear combinations of the elements in $\boldsymbol{y}_T$, which take the same *linear forms* as the corresponding VT point estimates. However, their results stop short of establishing numerical equivalence as in Theorem 1. From this perspective, Theorem 1 is perhaps surprising as it takes the next step forward in contradicting the notion that the two regressions are fundamentally different. In particular, Theorem 1 proves the same point estimate is derived from both approaches whenever the regression model belongs to the symmetric class, namely (i) OLS with minimum ℓ_2-norm, (ii) PCR for the same choice of k, and (iii) ridge for the same choice of λ_2. In this view, HZ and VT regressions are not perspectives at duel—they are dual perspectives.

We emphasize that Theorem 1 holds for any data configuration. As such, it clarifies that HZ and VT OLS are *not* invalid when $T > N$ and $N > T$, respectively, as previously believed. In fact, the OLS estimate can even be written as $\langle \widehat{\boldsymbol{\alpha}}, \boldsymbol{Y}_0' \widehat{\boldsymbol{\beta}} \rangle$, which incorporates both regression models. The origin of the prior misconception may have come from the fact that infinitely many solutions exist when $\boldsymbol{Y}_0$ is rank deficient. Among these solutions, however, is the unique minimum ℓ_2-norm model, which is arguably sufficient for inference (Shao and Deng, 2012). This is also the solution when the problem is optimized via gradient descent, a ubiquitous optimizer in practice. Phenomena of this form are known as "implicit regularization", where the optimization algorithm is biased towards a particular solution even though the bias is not explicit in the objective function (Neyshabur et al., 2015, Gunasekar et al., 2017).

Through its connection to the ℓ_2-penalty, the minimum ℓ_2-norm also offers a high-level intuition for the root of symmetry. More specifically, observe that the ridge model converges to the OLS model with minimum ℓ_2-norm as $\lambda_2 \to 0$. Since the PCR model is precisely the OLS minimum ℓ_2-norm model that is restricted to the space spanned by the top k principal components, we conjecture that the geometry of the ℓ_2-ball is a likely source for HZ and VT estimation symmetry.

II: Asymmetric class. Next, we state the class of formulations that fracture symmetry.

THEOREM 2: HZ $\neq$ VT *for (i) lasso, (ii) elastic net, and (iii) simplex regression.*

To examine the implications of Theorem 2, we observe that the common thread between the objective functions in the asymmetric class is a penalty or constraint that promotes sparse models. Such regularizers are noticeably absent in the symmetric formulations. This highlights an interesting trade-off—while sparsity is widely considered to be a salient feature, the geometries of the ℓ_1-ball and simplex that encourage sparsity are also the likely sources of HZ and VT estimation asymmetry.

3.2. *Doubly Robust Estimators*

In recent years, there has been a surge of interest in doubly robust estimators. Within panel data, we discuss two prominent works that are rising in popularity.

3.2.1. *Synthetic Difference-in-Differences*

An important approach that continues to dominate empirical work in panel data is the difference-in-differences (DID) estimator (Ashenfelter, 1978). At a high level, DID posits an additive outcome model with unit- and time-specific fixed effects, known more colloquially as the "parallel trends" assumption. The recent work of Arkhangelsky et al. (2021) anchors on the DID principle and brings in concepts from the unconfoundedness and synthetic controls literatures to derive a doubly robust estimator called synthetic difference-in-differences (SDID). In our setting, the SDID prediction for the missing (N, T)th potential outcome can be written as

$$\widehat{Y}_{NT}^{\text{sdid}}(0) = \sum_{i \leq N_0} \widehat{\beta}_i Y_{iT} + \sum_{t \leq T_0} \widehat{\alpha}_t Y_{Nt} - \sum_{i \leq N_0} \sum_{t \leq T_0} \widehat{\beta}_i \widehat{\alpha}_t Y_{it}$$

$$= \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle + \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle - \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{Y}_0 \widehat{\boldsymbol{\alpha}} \rangle, \tag{6}$$

where $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$ represent general HZ and VT models, respectively. The authors note that $\widehat{\boldsymbol{\alpha}} = (1/T_0)\mathbf{1}$ and $\widehat{\boldsymbol{\beta}} = (1/N_0)\mathbf{1}$ recovers DID. Moving beyond simple DID to performing a weighted two-way bias removal, the authors propose to learn $\widehat{\boldsymbol{\alpha}}$ via simplex regression and $\widehat{\boldsymbol{\beta}}$ via simplex regression with an ℓ_2-penalty.

3.2.2. *Augmented Synthetic Controls*

Another notable work is that of Ben-Michael et al. (2021). The authors introduce the augmented synthetic control (ASC) estimator, which uses an outcome model to correct the bias induced by the classical synthetic controls estimator.[1] Concretely, the ASC estimator predicts the missing (N, T)th potential outcome as

$$\widehat{Y}_{NT}^{\text{asc}}(0) = \widehat{M}_{NT}(0) + \sum_{i \leq N_0} \widehat{\beta}_i (Y_{iT} - \widehat{M}_{iT}(0)), \tag{7}$$

where $\widehat{M}_{iT}(0)$ is the estimator for the (i, T)th entry. The authors instantiate

$$\widehat{M}_{iT}(0) = \sum_{t \leq T_0} \widehat{\alpha}_t Y_{it}. \tag{8}$$

Plugging the HZ outcome model in (8) into (7) then gives

$$\widehat{Y}_{NT}^{\text{asc}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle + \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle - \langle \widehat{\boldsymbol{\beta}}, \boldsymbol{Y}_0 \widehat{\boldsymbol{\alpha}} \rangle.$$

We consider this particular variant of ASC since it takes the same form as SDID, as seen in (6). In contrast to Arkhangelsky et al. (2021), Ben-Michael et al. (2021) learns $\widehat{\boldsymbol{\alpha}}$ via ridge regression and $\widehat{\boldsymbol{\beta}}$ via simplex regression.

3.2.3. *When Doubly Robust Estimators are No Longer Doubly Robust*

We leverage Theorem 1 to study the estimation properties of SDID and ASC when $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ are learned via OLS and PCR.

---

[1]See Abadie and L'Hour (2021) for a bias correction of synthetic controls through matching.

COROLLARY 1: $\mathsf{SDID} = \mathsf{ASC} = \mathsf{HZ} = \mathsf{VT}$ *for (i) $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ as the OLS minimum ℓ_2-norm solutions and (ii) $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ as the PCR solutions with the same choice of $k < R$.*

Corollary 1 states that a researcher who uses the implicitly regularized (but explicitly unconstrained) variants of SDID and ASC inevitably arrive at the same point estimate as their colleague who simply uses HZ or VT OLS. The same phenomena occurs for PCR. From this, we observe that SDID and ASC can lose their weighted double-differencing effects under certain formulations. However, Corollary 1 is not to discredit either approach. The fact remains that both methods allow researchers to naturally and simultaneously encode their knowledge of time and unit specific structures into the estimator. SDID and ASC warrant further studies and careful consideration.

3.3. *Intercepts*

Intercepts can be included in the HZ regression model by modifying the ℓ_2-errors in (1) and (4) as $\|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha} - \alpha_0\boldsymbol{1}\|_2^2$; similarly, they can be included in the VT regression model by modifying ℓ_2-errors in (2) and (5) as $\|\boldsymbol{y}_N - \boldsymbol{Y}_0'\boldsymbol{\beta} - \beta_0\boldsymbol{1}\|_2^2$. To date, there is still a lack of general consensus on the inclusion of intercepts within the synthetic controls literature. We attempt to shed light on the role of intercepts.

COROLLARY 2: $\mathsf{HZ} \neq \mathsf{VT}$ *for (i) OLS, (ii) PCR, and (iii) ridge with intercepts.*

We develop an intuition for Proposition 2 by interpreting intercepts in panel studies. A nonzero time intercept, α_0, imposes a permanent constant difference between the treated and pretreatment periods; a nonzero unit intercept, β_0, imposes a permanent constant difference between the treated and control units. These systematic structures then create an asymmetry between the two regressions. Below, we propose a methodology based on centering the data that allows for intercepts yet retains symmetry.

3.3.1. *Including Intercepts and Retaining Symmetry through Data Centering*

Let $\boldsymbol{Y}_0$ be twice centered, i.e., the rows and columns of $\boldsymbol{Y}_0$ are mean zero. This can be satisfied by applying $\boldsymbol{I} - (1/N_0)\boldsymbol{1}\boldsymbol{1}'$ and $\boldsymbol{I} - (1/T_0)\boldsymbol{1}\boldsymbol{1}'$ to the left and right, respectively, of $\boldsymbol{Y}_0$. Consider the following modified formulations.

(a) HZ regression: for $\lambda \geq 0$,

$$(\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\boldsymbol{\alpha}}) = \underset{(\alpha_0, \alpha_1, \boldsymbol{\alpha})}{\operatorname{argmin}} \ \|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha} - \alpha_0\boldsymbol{1}\|_2^2 + \|\boldsymbol{y}_N - \alpha_1\boldsymbol{1}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2 \qquad (9)$$

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle + \widehat{\alpha}_0 + \widehat{\alpha}_1. \qquad (10)$$

(b) VT regression: for $\lambda \geq 0$,

$$(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\boldsymbol{\beta}}) = \underset{(\beta_0, \beta_1, \boldsymbol{\beta})}{\operatorname{argmin}} \ \|\boldsymbol{y}_N - \boldsymbol{Y}_0'\boldsymbol{\beta} - \beta_0\boldsymbol{1}\|_2^2 + \|\boldsymbol{y}_T - \beta_1\boldsymbol{1}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \qquad (11)$$

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle + \widehat{\beta}_0 + \widehat{\beta}_1. \qquad (12)$$

Similar to before, OLS corresponds to $\lambda = 0$, PCR corresponds to OLS with $\boldsymbol{Y}_0^{(k)}$ for $k < R$ in place of $\boldsymbol{Y}_0$, and ridge regression corresponds to any $\lambda > 0$.

COROLLARY 3: $\mathsf{HZ} = \mathsf{VT}$ *for the symmetric estimators in Theorem 1 under the formulations set in* (9) *and* (11) *with* $\boldsymbol{Y}_0$ *being twice centered.*

We inspect (10) and (12) to understand the implications of Corollary 3. First, we recall Theorem 1, which establishes that the HZ and VT estimates share the same "base" estimate, i.e., $\langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle$. Next, we note that $\widehat{\alpha}_0 = \widehat{\beta}_1 = (1/N_0)\mathbf{1}'\boldsymbol{y}_T$ and $\widehat{\beta}_0 = \widehat{\alpha}_1 = (1/T_0)\mathbf{1}'\boldsymbol{y}_N$, which correspond to the time and unit fixed effects, respectively. Intuitively, the modified point estimates in (10) and (12) include both fixed effect models to compensate for $\boldsymbol{Y}_0$ being twice centered. Putting everything together, the modified HZ and VT point estimates are identical.

4. INFERENCE

> **Q2:** "When the HZ and VT point estimates are identical, how does the source of randomness impact inference?"

To answer Q2, we study the inferential properties of the counterfactual prediction. Formal discussions for classical inference require an explicit postulation of where the randomness arises from. This article takes both a (i) model-based approach, which makes assumptions about the distribution of the potential outcomes, and a (ii) design-based approach, which makes assumptions about the assignment mechanism of treatment. We start by taking a model-based route that is similar in parts to the recent works of Li (2020), Cattaneo et al. (2021), and Chernozhukov et al. (2021, 2022) from the synthetic controls literature. We then transition to a design-based route that is paved by the ideas introduced in Bottmer et al. (2021).

For ease of exposition, we focus on OLS and its minimum ℓ_2-norm solutions, i.e., $\widehat{\boldsymbol{\alpha}} = \boldsymbol{Y}_0^\dagger \boldsymbol{y}_T$ and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}_0')^\dagger \boldsymbol{y}_N$. As such, we define $\widehat{Y}_{NT}(0) = \widehat{Y}_{NT}^{\mathrm{hz}}(0) = \widehat{Y}_{NT}^{\mathrm{vt}}(0)$, which is justified under Theorem 1. To improve readability, we present several results informally and provide their precise versions in Appendix A.

4.1. *Model-Based Inference*

Within the model-based framework, we consider a classical regression model. This postulation is not always plausible but it is useful to dissect its implications for the role of randomness in conducting inference.

4.1.1. *Generative Models*

We now study properties of $\widehat{Y}_{NT}(0)$ from three different sources of randomness.

I: Horizontal model. The HZ model focuses on time series correlation patterns.

ASSUMPTION 1: *Conditional on* $(\boldsymbol{y}_N, \boldsymbol{Y}_0)$, *we have*

$$Y_{iT} = \sum_{t \leq T_0} \alpha_t^* Y_{it} + \varepsilon_{iT}, \quad i = 1, \ldots, N_0,$$

where $\boldsymbol{\alpha}^*$ *is a vector of unknown coefficients and* ε_{iT} *is an idiosyncratic error term.*

Assumption 1 motivates the HZ approach, which models time series patterns as the source of randomness. Under the HZ model, the statistical uncertainty of $\widehat{Y}_{NT}(0)$ is governed by the construction of $\widehat{\boldsymbol{\alpha}}$ from $(\boldsymbol{y}_T, \boldsymbol{Y}_0)$, i.e., the in-sample uncertainty.

ASSUMPTION 2: $\{\varepsilon_{iT}\}_{i=1}^{N_0}$ has zero mean and is independent over $i = 1, \ldots, N_0$, conditional on $(\boldsymbol{y}_N, \boldsymbol{Y}_0)$.

Assumption 2 states that the errors have zero mean and thus the regressors, i.e., lagged outcomes, are uncorrelated with the errors; this is known in the literature as strict exogeneity. The errors are also conditionally independent across space.

II: Vertical model. The VT model focuses on cross-sectional correlation patterns.

ASSUMPTION 3: Conditional on $(\boldsymbol{y}_T, \boldsymbol{Y}_0)$, we have

$$Y_{Nt} = \sum_{i \le N_0} \beta_i^* Y_{it} + \varepsilon_{Nt}, \quad t = 1, \ldots, T_0,$$

where $\boldsymbol{\beta}^*$ is a vector of unknown coefficients and ε_{Nt} is an idiosyncratic error term.

Assumption 3 is analogous to Assumption 1 with the source of randomness here emanating from cross-sectional patterns. Hence, the statistical uncertainty of $\widehat{Y}_{NT}(0)$ under the VT model is governed by the construction of $\widehat{\boldsymbol{\beta}}$ from $(\boldsymbol{y}_N, \boldsymbol{Y}_0)$.

ASSUMPTION 4: $\{\varepsilon_{Nt}\}_{t=1}^{T_0}$ has zero mean and is independent over $t = 1, \ldots, T_0$, conditional on $(\boldsymbol{y}_T, \boldsymbol{Y}_0)$.

Assumption 4 is analogous to Assumption 2 with the errors here being conditionally independent across time.

III: Mixed model. We introduce a new model that mixes aspects of the HZ and VT models. At a high level, the mixed model accounts for randomness along both dimensions of the data. This comes at the price of placing additional constraints on the stochastic properties of the errors.

ASSUMPTION 5: Conditional on $\boldsymbol{Y}_0$, we have

$$Y_{iT} = \sum_{t \le T_0} \alpha_t^* Y_{it} + \varepsilon_{iT}, \quad i = 1, \ldots, N_0, \tag{13}$$

$$Y_{Nt} = \sum_{i \le N_0} \beta_i^* Y_{it} + \varepsilon_{Nt}, \quad t = 1, \ldots, T_0, \tag{14}$$

where $(\boldsymbol{\alpha}^*, \varepsilon_{iT})$ and $(\boldsymbol{\beta}^*, \varepsilon_{Nt})$ are defined as in Assumptions 1 and 3, respectively.

(13) and (14) correspond to Assumptions 1 and 3, respectively. Collectively, they model time series and cross-sectional patterns as two distinct sources of randomness. Thus, the statistical uncertainty of $\widehat{Y}_{NT}(0)$ under the mixed model is governed by the constructions of both $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$.

ASSUMPTION 6: $\{\varepsilon_{iT}\}_{i=1}^{N_0}$ and $\{\varepsilon_{Nt}\}_{t=1}^{T_0}$ have zero mean and are independent over $i = 1, \ldots, N_0$ and $t = 1, \ldots, T_0$, conditional on $\boldsymbol{Y}_0$.

Assumption 6 combines Assumptions 2 and 4. In words, it states that $\boldsymbol{Y}_0$ contains all measured confounders and the errors are independent across both time and space. We leave a proper analysis under dependent errors as future work.

4.1.2. *Inferential Properties*

Equipped with our three models, we are ready to provide one set of answers to Q2. In what follows, we define the error covariance matrices as (i) $\boldsymbol{\Sigma}_T^{\text{hz}} = \text{Cov}(\boldsymbol{\varepsilon}_T | \boldsymbol{y}_N, \boldsymbol{Y}_0)$, (ii) $\boldsymbol{\Sigma}_N^{\text{vt}} = \text{Cov}(\boldsymbol{\varepsilon}_N | \boldsymbol{y}_T, \boldsymbol{Y}_0)$, (iii) $\boldsymbol{\Sigma}_T^{\text{mix}} = \text{Cov}(\boldsymbol{\varepsilon}_T | \boldsymbol{Y}_0)$, and (iv) $\boldsymbol{\Sigma}_N^{\text{mix}} = \text{Cov}(\boldsymbol{\varepsilon}_N | \boldsymbol{Y}_0)$, where $\boldsymbol{\varepsilon}_T = [\varepsilon_{iT} : i \leq N_0]$ and $\boldsymbol{\varepsilon}_N = [\varepsilon_{Nt} : t \leq T_0]$.

THEOREM 3: *(i)* [HZ model] *Under Assumptions 1–2 and suitable moment conditions, we have*

$$\frac{\widehat{Y}_{NT}(0) - \mu_0^{\text{hz}}}{\sqrt{v_0^{\text{hz}}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mu_0^{\text{hz}} = \langle \boldsymbol{y}_N, \boldsymbol{H}^v \boldsymbol{\alpha}^* \rangle$ *and* $v_0^{\text{hz}} = \widehat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_T^{\text{hz}} \widehat{\boldsymbol{\beta}}$. *(ii)* [VT model] *Under Assumptions 3–4 and suitable moment conditions, we have*

$$\frac{\widehat{Y}_{NT}(0) - \mu_0^{\text{vt}}}{\sqrt{v_0^{\text{vt}}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mu_0^{\text{vt}} = \langle \boldsymbol{y}_T, \boldsymbol{H}^u \boldsymbol{\beta}^* \rangle$ *and* $v_0^{\text{vt}} = \widehat{\boldsymbol{\alpha}}' \boldsymbol{\Sigma}_N^{\text{vt}} \widehat{\boldsymbol{\alpha}}$. *(iii)* [Mixed model] *Under Assumptions 5–6 and suitable moment conditions, we have*

$$\frac{\widehat{Y}_{NT}(0) - \mu_0^{\text{mix}}}{\sqrt{v_0^{\text{mix}}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mu_0^{\text{mix}} = \langle \boldsymbol{\alpha}^*, \boldsymbol{Y}_0' \boldsymbol{\beta}^* \rangle$ *and*

$$v_0^{\text{mix}} = (\boldsymbol{H}^u \boldsymbol{\beta}^*)' \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{H}^u \boldsymbol{\beta}^*) + (\boldsymbol{H}^v \boldsymbol{\alpha}^*)' \boldsymbol{\Sigma}_N^{\text{mix}} (\boldsymbol{H}^v \boldsymbol{\alpha}^*) + \text{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\text{mix}}).$$

Theorem 3 highlights that each model measures uncertainty with respect to a different estimand. In particular, Theorem 3 states that the asymptotic variance is controlled by time series patterns under the HZ model, cross-sectional patterns under the VT model, and both correlation patterns under the mixed model. This clarifies that the source of randomness has substantive implications for the estimand and inference.

4.1.3. *Confidence Intervals*

Theorem 3 motivates separate HZ, VT, and mixed confidence intervals: for $\theta \in (0, 1)$,

$$\mu_0^{\text{hz}} \in \left[\widehat{Y}_{NT}(0) \ \pm \ z_{\frac{\theta}{2}} \sqrt{\widehat{v_0^{\text{hz}}}} \right],$$

$$\mu_0^{\text{vt}} \in \left[\widehat{Y}_{NT}(0) \ \pm \ z_{\frac{\theta}{2}} \sqrt{\widehat{v_0^{\text{vt}}}} \right],$$

$$\mu_0^{\text{mix}} \in \left[\widehat{Y}_{NT}(0) \ \pm \ z_{\frac{\theta}{2}} \sqrt{\widehat{v_0^{\text{mix}}}} \right],$$

where $z_{\frac{\theta}{2}}$ is the upper $\theta/2$ quantile of $\mathcal{N}(0,1)$, and $(\widehat{v}_0^{\mathrm{hz}}, \widehat{v}_0^{\mathrm{vt}}, \widehat{v}_0^{\mathrm{mix}})$ are the estimators of $(v_0^{\mathrm{hz}}, v_0^{\mathrm{vt}}, v_0^{\mathrm{mix}})$. We construct

$$\widehat{v}_0^{\mathrm{hz}} = \widehat{\boldsymbol{\beta}}' \widehat{\boldsymbol{\Sigma}}_T \widehat{\boldsymbol{\beta}}, \quad \widehat{v}_0^{\mathrm{vt}} = \widehat{\boldsymbol{\alpha}}' \widehat{\boldsymbol{\Sigma}}_N \widehat{\boldsymbol{\alpha}}, \quad \widehat{v}_0^{\mathrm{mix}} = \widehat{v}_0^{\mathrm{hz}} + \widehat{v}_0^{\mathrm{vt}} - \mathrm{tr}(\boldsymbol{Y}_0^{\dagger} \widehat{\boldsymbol{\Sigma}}_T (\boldsymbol{Y}_0')^{\dagger} \widehat{\boldsymbol{\Sigma}}_N), \qquad (15)$$

where $\widehat{\boldsymbol{\Sigma}}_T$ and $\widehat{\boldsymbol{\Sigma}}_N$ are the estimators of $(\boldsymbol{\Sigma}_T^{\mathrm{hz}}, \boldsymbol{\Sigma}_T^{\mathrm{mix}})$ and $(\boldsymbol{\Sigma}_N^{\mathrm{vt}}, \boldsymbol{\Sigma}_N^{\mathrm{mix}})$, respectively. We precisely define them under homoskedastic and heteroskedastic errors below. To reduce ambiguity, we index $(\widehat{v}_0^{\mathrm{hz}}, \widehat{v}_0^{\mathrm{vt}}, \widehat{v}_0^{\mathrm{mix}})$ by the covariance estimator. We also denote $\boldsymbol{H}^u = \boldsymbol{U}\boldsymbol{U}'$ and $\boldsymbol{H}_{\perp}^u = \boldsymbol{I} - \boldsymbol{H}^u$. With this notation, the HZ and VT in-sample errors can be written as $\boldsymbol{H}_{\perp}^u \boldsymbol{y}_T = \boldsymbol{y}_T - \boldsymbol{Y}_0 \widehat{\boldsymbol{\alpha}}$ and $\boldsymbol{H}_{\perp}^v \boldsymbol{y}_N = \boldsymbol{y}_N - \boldsymbol{Y}_0' \widehat{\boldsymbol{\beta}}$, respectively.

It is clear from (15) that $(\widehat{v}_0^{\mathrm{hz}}, \widehat{v}_0^{\mathrm{vt}})$ are plug-in estimators for $(v_0^{\mathrm{hz}}, v_0^{\mathrm{vt}})$. As such, we discuss $\widehat{v}_0^{\mathrm{mix}}$ with respect to v_0^{mix}. Recall $\widehat{\boldsymbol{\alpha}} = \boldsymbol{H}^v \widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}} = \boldsymbol{H}^u \widehat{\boldsymbol{\beta}}$ by construction. To justify the negative trace in $\widehat{v}_0^{\mathrm{mix}}$, note that $\widehat{v}_0^{\mathrm{hz}}$ is a quadratic involving $(\boldsymbol{y}_N, \boldsymbol{y}_T)$. Since both quantities are random, the expectation of $\widehat{v}_0^{\mathrm{hz}}$ induces an additional term that precisely corresponds to the trace term in v_0^{mix}. The same property holds for $\widehat{v}_0^{\mathrm{vt}}$. Thus, $\widehat{v}_0^{\mathrm{mix}}$ corrects for this bias via the negative trace.

Homoskedastic errors. Consider $\boldsymbol{\Sigma}_T^{\mathrm{hz}}$ with identical diagonal elements, i.e., $\boldsymbol{\Sigma}_T^{\mathrm{hz}} = (\sigma_T^{\mathrm{hz}})^2 \boldsymbol{I}$, where $(\sigma_T^{\mathrm{hz}})^2 = \mathrm{Var}(\varepsilon_{iT}|\boldsymbol{y}_N, \boldsymbol{Y}_0)$ for $i = 1, \ldots, N_0$. Let $(\boldsymbol{\Sigma}_N^{\mathrm{vt}}, \boldsymbol{\Sigma}_T^{\mathrm{mix}}, \boldsymbol{\Sigma}_N^{\mathrm{mix}})$ be defined analogously. We use the standard variance estimators

$$\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{homo}} = \frac{1}{N_0 - R} \| \boldsymbol{H}_{\perp}^u \boldsymbol{y}_T \|_2^2 \boldsymbol{I}, \qquad (16)$$

$$\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{homo}} = \frac{1}{T_0 - R} \| \boldsymbol{H}_{\perp}^v \boldsymbol{y}_N \|_2^2 \boldsymbol{I}, \qquad (17)$$

where $R = \mathrm{rank}(\boldsymbol{Y}_0)$, which can be computed as $R = \mathrm{tr}(\boldsymbol{H}^u) = \mathrm{tr}(\boldsymbol{H}^v)$.

LEMMA 1: *Consider homoskedastic errors. (i)* [HZ model] *Under Assumptions 1–2, we have*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{homo}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\mathrm{hz}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\mathrm{hz,homo}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = v_0^{\mathrm{hz}}.$$

(ii) [VT model] *Under Assumptions 3–4, we have*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{homo}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{vt}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\mathrm{vt,homo}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = v_0^{\mathrm{vt}}.$$

(iii) [Mixed model] *Under Assumptions 5–6, we have*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{homo}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\mathrm{mix}}, \quad \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{homo}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{mix}}, \quad and \quad \mathbb{E}[\widehat{v}_0^{\mathrm{mix,homo}}|\boldsymbol{Y}_0] = v_0^{\mathrm{mix}}.$$

Lemma 1 is a well known result within the OLS literature, albeit it is typically formalized under the stricter full column rank assumption.

Heteroskedastic errors. We adopt two strategies for the heteroskedastic setting.

I: Jackknife. The first estimator is based on the jackknife. Traditionally, the jackknife estimates the covariance of the regression coefficients $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$. By analyzing said estimates, we derive the following:

$$\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{jack}} = \mathrm{diag}\left([\boldsymbol{H}_{\perp}^u \circ \boldsymbol{H}_{\perp}^u \circ \boldsymbol{I}]^{\dagger} [\boldsymbol{H}_{\perp}^u \boldsymbol{y}_T \circ \boldsymbol{H}_{\perp}^u \boldsymbol{y}_T] \right) \qquad (18)$$

$$\widehat{\boldsymbol{\Sigma}}_N^{\text{jack}} = \text{diag}\left([\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v \circ \boldsymbol{I}]^\dagger [\boldsymbol{H}_\perp^v \boldsymbol{y}_N \circ \boldsymbol{H}_\perp^v \boldsymbol{y}_N]\right). \tag{19}$$

LEMMA 2: *Consider heteroskedastic errors. (i)* [HZ model] *Let Assumptions 1–2 hold. If* $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})$ *is nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}}|\boldsymbol{y}_N,\boldsymbol{Y}_0] \succeq \boldsymbol{\Sigma}_T^{\text{hz}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{hz,jack}}|\boldsymbol{y}_N,\boldsymbol{Y}_0] \geq v_0^{\text{hz}}.$$

(ii) [VT model] *Let Assumptions 3–4 hold. If* $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v \circ \boldsymbol{I})$ *is nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{jack}}|\boldsymbol{y}_T,\boldsymbol{Y}_0] \succeq \boldsymbol{\Sigma}_N^{\text{vt}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{vt,jack}}|\boldsymbol{y}_T,\boldsymbol{Y}_0] \geq v_0^{\text{vt}}.$$

(iii) [Mixed model] *Let Assumptions 5–6 hold. If* $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})$ *and* $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v \circ \boldsymbol{I})$ *are nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}}|\boldsymbol{Y}_0] \succeq \boldsymbol{\Sigma}_T^{\text{mix}}, \quad \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{jack}}|\boldsymbol{Y}_0] \succeq \boldsymbol{\Sigma}_N^{\text{mix}}, \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{mix,jack}}|\boldsymbol{Y}_0] \geq v_0^{\text{mix}}.$$

Lemma 2 establishes that the jackknife is conservative, provided $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})$ and $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v \circ \boldsymbol{I})$ are nonsingular. Strictly speaking, the jackknife is well defined if these quantities are singular, as seen through the pseudoinverse in (18) and (19). Lemma 2 considers the nonsingular case for simplicity. We remark that $\max_\ell H_{\ell\ell}^u < 1$ and $\max_\ell H_{\ell\ell}^v < 1$ are sufficient conditions for invertibility.

II: HRK-estimator. Next, we consider the covariance estimator proposed by Hartley et al. (1969). We index this estimator by the authors, Hartley-Rao-Kiefer:

$$\widehat{\boldsymbol{\Sigma}}_T^{\text{HRK}} = \text{diag}\left([\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u]^{-1}[\boldsymbol{H}_\perp^u \boldsymbol{y}_T \circ \boldsymbol{H}_\perp^u \boldsymbol{y}_T]\right)$$

$$\widehat{\boldsymbol{\Sigma}}_N^{\text{HRK}} = \text{diag}\left([\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v]^{-1}[\boldsymbol{H}_\perp^v \boldsymbol{y}_N \circ \boldsymbol{H}_\perp^v \boldsymbol{y}_N]\right).$$

LEMMA 3: *Consider heteroskedastic errors. (i)* [HZ model] *Let Assumptions 1–2 hold. If* $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)$ *is nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{HRK}}|\boldsymbol{y}_N,\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\text{hz}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{hz,HRK}}|\boldsymbol{y}_N,\boldsymbol{Y}_0] = v_0^{\text{hz}}.$$

(ii) [VT model] *Let Assumptions 3–4 hold. If* $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v)$ *is nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{HRK}}|\boldsymbol{y}_T,\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\text{vt}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{vt,HRK}}|\boldsymbol{y}_T,\boldsymbol{Y}_0] = v_0^{\text{vt}}.$$

(iii) [Mixed model] *Let Assumptions 5–6 hold. If* $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)$ *and* $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v)$ *are nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{HRK}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\text{mix}}, \quad \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{HRK}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\text{mix}}, \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{mix,HRK}}|\boldsymbol{Y}_0] = v_0^{\text{mix}}.$$

Lemma 3 establishes that the HRK estimator is unbiased, provided $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)$ and $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v)$ are invertible. For the former quantity, $\max_\ell H_{\ell\ell}^u < 1/2$ is a sufficient condition for invertibility. Since $\text{tr}(\boldsymbol{H}^u) = R$, this restricts $R < N_0/2$. A similar conclusion is drawn for VT regression.

<div align="center">4.2. Design-Based Inference</div>

This section studies the counterfactual prediction from a design-based perspective, whereby the potential outcomes are considered fixed and the treatment assignments are considered stochastic.

4.2.1. *Assumptions*

In congruence with the article thus far, we focus on a single treated unit and treated period. We will find it useful to separate the assignment mechanism into the selection of each component. Accordingly, let $\boldsymbol{A} \in \{0,1\}^T$ with $\boldsymbol{1}'\boldsymbol{A} = 1$ and $\boldsymbol{B} \in \{0,1\}^N$ with $\boldsymbol{1}'\boldsymbol{B} = 1$ be the indicator vectors for the treated time period and treated unit, respectively, e.g., $B_N = 1$ and $A_T = 1$ if unit N is treated at time T. With this notation, we denote the realized outcome as $Y_{it} = B_i A_t Y_{it}(1) + (1 - B_i A_t) Y_{it}(0)$. Following Bottmer et al. (2021), we consider the following assignment mechanisms:

ASSUMPTION 7—Random assignment of time period:

$$\mathbb{P}(\boldsymbol{A} = \boldsymbol{a}) = \left\{ \begin{array}{ll} 1/T, & \textit{if } a_t \in \{0,1\} \ \forall t, \ \boldsymbol{1}'\boldsymbol{a} = 1 \\ 0, & \textit{otherwise.} \end{array} \right.$$

ASSUMPTION 8—Random assignment of unit:

$$\mathbb{P}(\boldsymbol{B} = \boldsymbol{b}) = \left\{ \begin{array}{ll} 1/N, & \textit{if } b_i \in \{0,1\} \ \forall i, \ \boldsymbol{1}'\boldsymbol{b} = 1 \\ 0, & \textit{otherwise.} \end{array} \right.$$

Assumption 7 considers the treated period to be randomly selected while Assumption 8 considers the treated unit to be randomly selected. As Bottmer et al. (2021) notes, these assumptions are not always plausible, but they underlie the placebo tests that are commonly used in synthetic controls applications.

4.2.2. *Estimator*

To conduct design-based analysis, we consider all possible treatment assignments, not only the realized assignment. Let $Y_{it}^*(0)$ be the OLS fit of $Y_{it}(0)$ based on $[Y_{j\tau} : j \neq i, \tau < t]$, $[Y_{i\tau} : \tau < t]$, and $[Y_{jt} : j \neq i]$. We define the design-based estimator as

$$\widehat{Y}(0) = \sum_{i \leq N} \sum_{t \leq T} B_i A_t Y_{it}^*(0). \tag{20}$$

In words, (20) predicts the mean counterfactual outcome under control for unit i at time t if unit i is treated at time period t, i.e., $B_i = 1$ and $A_t = 1$. We reemphasize that the stochasticity of $\widehat{Y}(0)$ stems from the treatment assignment mechanism since $Y_{it}^*(0)$ is a fixed quantity. As such, while the model-and design-based estimators share the same point estimate for the realized assignment, they differ in their formulations and attributions of randomness.

4.2.3. *Inferential Properties*

In Table I, we summarize the estimands associated with the model-based and design-based estimators under three sources of randomness: (i) time, (ii) unit, and (iii) time

and unit. Within the model-based framework, mechanisms (i)–(iii) correspond to the HZ model (Assumptions 1–2), VT model (Assumptions 3–4), and mixed model (Assumptions 5–6), respectively. Within the design-based framework, mechanisms (i)–(iii) correspond to Assumption 7, Assumption 8, and both Assumptions 7 and 8, respectively. Accordingly, the model-based and design-based expectations are taken over different probability measures.

TABLE I

MODEL-BASED AND DESIGN-BASED ESTIMANDS UNDER DIFFERENT SOURCES OF RANDOMNESS. WE USE THE SHORTHAND $\tilde{\boldsymbol{\alpha}}^* = \boldsymbol{H}^v \boldsymbol{\alpha}^*$ AND $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{H}^u \boldsymbol{\beta}^*$ AND CONSIDER THE REALIZED (N,T)TH ASSIGNMENT.

Source of Randomness	Model-Based Estimand	Design-Based Estimand
Time	$\mathbb{E}[\widehat{Y}_{NT}(0)\|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \sum_{t \leq T_0} \tilde{\alpha}_t^* Y_{Nt}$	$\mathbb{E}[\widehat{Y}(0)\|\boldsymbol{B}] = \frac{1}{T} \sum_{t \leq T} Y_{Nt}^*(0)$
Unit	$\mathbb{E}[\widehat{Y}_{NT}(0)\|\boldsymbol{y}_T, \boldsymbol{Y}_0] = \sum_{i \leq N_0} \tilde{\beta}_i^* Y_{iT}$	$\mathbb{E}[\widehat{Y}(0)\|\boldsymbol{A}] = \frac{1}{N} \sum_{i \leq N} Y_{iT}^*(0)$
Time and Unit	$\mathbb{E}[\widehat{Y}_{NT}(0)\|\boldsymbol{Y}_0] = \sum_{i \leq N_0} \sum_{t \leq T_0} \alpha_t^* \beta_i^* Y_{it}$	$\mathbb{E}[\widehat{Y}(0)] = \frac{1}{NT} \sum_{i \leq N} \sum_{t \leq T} Y_{it}^*(0)$

Let us compare the estimands in Table I. In words, $\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0]$ is a weighted combination of outcomes under control for the treated unit across all T_0 pretreatment periods; $\mathbb{E}[\widehat{Y}(0)|\boldsymbol{B}]$ is a simple average of fitted outcomes under control for the treated unit across all T periods, which Bottmer et al. (2021) calls the "HZ" effect. Similarly, $\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{y}_T, \boldsymbol{Y}_0]$ is a weighted combination of outcomes under control for the treated period across all N_0 control units; $\mathbb{E}[\widehat{Y}(0)|\boldsymbol{A}]$ is a simple average of fitted outcomes under control for the treated period across all N units, also called the "VT" effect. Finally, $\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0]$ is a weighted combination of outcomes under control across all N_0 control units and T_0 pretreatment periods; $\mathbb{E}[\widehat{Y}(0)]$ is a simple average of fitted outcomes under control across all unit and time period pairs, which we coin the "mixed" effect.

Though our design-based analysis is brief, there two important takeaways: (i) the model-based and design-based estimators recover similar estimands for each source of randomness; and (ii) different sources of randomness lead to different estimands, which is consistent with our model-based insights. The connection between assumptions of randomness and resulting estimand has been previously noticed in related contexts, e.g., Abadie et al. (2020), Bottmer et al. (2021) and Sekhon and Shem-Tov (2021).

4.3. *Discussion*

Several remarks on this section's results and extensions are in order.

REMARK 1—Correct Specification: For expositional convenience, we assume correct specification of the outcome model, as in Li (2020), to explain the panel data intuition and main theoretical results in a simple and transparent fashion. Alternatively, we can interpret Assumptions 1, 3, and 5 as linear prediction models that are detached from structural meanings a la Cattaneo et al. (2021) and Chernozhukov et al. (2021, 2022). For instance, within the mixed framework, we can redefine $\boldsymbol{\alpha}^* = \arg\min \mathbb{E}[\|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha}\|_2^2|\boldsymbol{Y}_0]$ as the best linear approximation of $\boldsymbol{y}_T$ based on $\boldsymbol{Y}_0$, conditional on $\boldsymbol{Y}_0$. Such models can be justified via factor models and vector autoregressive models (Chernozhukov et al., 2022).

REMARK 2—Mixed Model: Any estimator, whether it falls within the symmetric or asymmetric class, can be studied under the mixed model (Assumptions 5–6). There are also numerous ways to encode the mixed perspective beyond our postulation, e.g., random assignment of periods and units (Assumptions 7–8) of Bottmer et al. (2021).

REMARK 3—Mixed Variance Estimators: We highlight that Lemmas 1–3 only hold in expectation. For any particular realization, $\widehat{v}_0^{\text{mix}}$ may exhibit unexpected properties. For instance, if $\text{tr}(\boldsymbol{Y}_0^\dagger \widehat{\boldsymbol{\Sigma}}_T (\boldsymbol{Y}_0')^\dagger \widehat{\boldsymbol{\Sigma}}_N) > \max\{\widehat{v}_0^{\text{hz}}, \widehat{v}_0^{\text{vt}}\}$, then $\widehat{v}_0^{\text{mix}} < \min\{\widehat{v}_0^{\text{hz}}, \widehat{v}_0^{\text{vt}}\}$; thus, the mixed coverage will be smaller than both HZ and VT coverages. In fact, $\widehat{v}_0^{\text{mix}}$ can be negative if $\text{tr}(\boldsymbol{Y}_0^\dagger \widehat{\boldsymbol{\Sigma}}_T (\boldsymbol{Y}_0')^\dagger \widehat{\boldsymbol{\Sigma}}_N) > \widehat{v}_0^{\text{hz}} + \widehat{v}_0^{\text{vt}}$, which may occur if both HZ and VT in-sample errors are "too large". For these scenarios, one naïve solution is to modify $\widehat{v}_0^{\text{mix}}$ as $\widehat{v}_0^{\text{mix}} \leftarrow \widehat{v}_0^{\text{hz}} + \widehat{v}_0^{\text{vt}}$, which is conservative by Lemmas 1–3. However, this case is arguably better resolved with a different point estimator altogether.

REMARK 4—Bridging Model-Based and Design-Based Inferences: The takeaways from the model-based and design-based analyses are consistent with one another. However, a formal and complete connection between the two perspectives on inference remains to be established. Towards this, we highlight that Lin (2013) demonstrates that one set of model-based confidence intervals (i.e., the Huber-White sandwich estimator) are justifiable under design-based arguments. In this view, a fascinating line of future inquiry is to analyze whether similar arguments hold for the heteroskedastic confidence intervals presented in Section 4.1.3 or other related model-based constructions.

REMARK 5—On the Role of Randomness: This section underscores that the source of randomness plays a pivotal role for conducting inference. Translated to practice, our results stress that researchers should be scrupulous in reasoning through where the randomness in their data comes from. For instance, in the Basque study, some researchers may find it more plausible that the randomness is over time rather than over space, e.g., it is more conceivable that the onset of terrorism could have occurred in a different year but less conceivable that it could have occurred in a different region of Spain but not in the Basque Country. We do not take a substantive view on the matter, but we do stress that these decisions are meaningful for proper analysis as they have immediate implications for the resulting estimand and inferential procedure.

REMARK 6—Model Checking: It is possible in an application that substantive knowledge does not make clear on the source of randomness and hence which model to use, i.e., HZ, VT, or mixed. In such scenarios, the "in-space" and "in-time" placebo tests (a la cross-validation) proposed in Abadie et al. (2010, 2015) are attractive tools to analyze the prediction properties of the various estimators under consideration. More generally, practices established within Yu and Kumbier (2020) provide an organized framework based on the principles of predictability, computability, and stability (PCS) to conduct rigorous comparison analyses.

REMARK 7—Extension to PCR: The previous results immediately extend to PCR by replacing $\boldsymbol{Y}_0$ with $\boldsymbol{Y}_0^{(k)}$, as defined in (3), for any $k < R$. Intuitively, PCR-based models operate under the belief that the data is inherently low-dimensional. We comment on several benefits of PCR over OLS. To begin, the HZ and VT OLS variance estimators constructed in Section 4.1.3 can suffer from degeneracy when N and T are of different sizes. That is, if $N < T$, then the HZ in-sample error is likely zero (otherwise known as

overfitting), which causes the HZ coverage to collapse on the point estimate; analogous statements hold for the VT coverage when $N > T$. The PCR-based variance estimators, on the other hand, can avoid degeneracy through the number of chosen principal components k (regularization). On a related note, the nonsingularity conditions required for the jackknife and HRK variance estimators can also be by controlled by k.

5. ILLUSTRATIONS

This section illustrates key concepts developed in this article. Our report is based on three canonical synthetic controls studies: (i) terrorism in Basque Country, (ii) California's Proposition 99 (Abadie et al., 2010), and (iii) the reunification of West Germany (Abadie et al., 2015). In particular, we will conduct a model-based analysis using the confidence intervals developed in Section 4.1.3. We provide an overview of the results and relegate details (e.g., implementation) to Appendix B.

5.1. *Background on Case Studies*

Basque study. See Sections 1–2 for details.

California study. This study examines the effect of California's Proposition 99, an anti-tobacco legislation, on its tobacco consumption. The panel data contains per capita cigarette sales of $N = 39$ U.S. states over $T = 31$ years. There are $T_0 = 18$ pretreatment observations and $N_0 = 38$ control units. Our interest is to estimate California's cigarette sales in the absence of Proposition 99.

West Germany study. This study examines the economic impact of the 1990 reunification in West Germany. The panel data contains per capita GDP of $N = 17$ countries over $T = 44$ years. There are $T_0 = 30$ pretreatment observations and $N_0 = 16$ control units. Our interest is to estimate West Germany's GDP in the absence of reunification.

5.2. *Data-Inspired Simulation Studies*

We look to better understand the trade-offs in conducting inference under different sources of randomness. In an attempt to document our analysis in a realistic environment, we calibrate our simulations to our three studies.

5.2.1. *Data Generating Process*

We consider the single treated unit and time period setting. Specifically, we consider the actual treated unit, e.g., Basque Country, and focus on the first post-treatment period, e.g., one year after the outset of terrorism; hence, $T \leftarrow T_0 + 1$. Using the actual data, we generate the underlying regression models as

$$\boldsymbol{\alpha}^* = \operatorname*{argmin}_{\boldsymbol{\alpha}} \|\boldsymbol{y}_T^* - \boldsymbol{Y}_0^* \boldsymbol{\alpha}\|_2^2 \quad \text{and} \quad \boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y}_N^* - (\boldsymbol{Y}_0^*)'\boldsymbol{\beta}\|_2^2,$$

where $\boldsymbol{y}_N^* = [Y_{Nt} : t \leq T_0]$, $\boldsymbol{y}_T^* = [Y_{iT} : i \leq N_0]$, and $\boldsymbol{Y}_0^* = [Y_{it} : i \leq N_0, t \leq T_0]$.

Observationally, we have access to the following quantities. Let $\boldsymbol{Y}_0$ be the rank r approximation of $\boldsymbol{Y}_0^*$, where r is chosen as the minimum number of singular values needed to capture at least 99.9% of $\boldsymbol{Y}_0^*$'s spectral energy. Next, we sample $\boldsymbol{y}_T \sim \mathcal{N}(\boldsymbol{Y}_0 \boldsymbol{\alpha}^*, (N_0 - r)^{-1}\|\boldsymbol{y}_T^* - \boldsymbol{Y}_0^*\boldsymbol{\alpha}^*\|_2^2 \boldsymbol{I})$ and $\boldsymbol{y}_N \sim \mathcal{N}(\boldsymbol{Y}_0'\boldsymbol{\beta}^*, (T_0 - r)^{-1}\|\boldsymbol{y}_N^* - (\boldsymbol{Y}_0^*)'\boldsymbol{\beta}^*\|_2^2 \boldsymbol{I})$. We then define three estimands: (i) $\mu_0^{\text{hz}} = \langle \boldsymbol{y}_N, \boldsymbol{H}^v \boldsymbol{\alpha}^* \rangle$, (ii) $\mu_0^{\text{vt}} = \langle \boldsymbol{y}_T, \boldsymbol{H}^u \boldsymbol{\beta}^* \rangle$, and (iii) $\mu_0^{\text{mix}} = \langle \boldsymbol{\alpha}^*, \boldsymbol{Y}_0'\boldsymbol{\beta}^* \rangle$, where $(\boldsymbol{H}^u, \boldsymbol{H}^v)$ are computed from $\boldsymbol{Y}_0$.

TABLE II

COVERAGE FOR NOMINAL 95% CONFIDENCE INTERVALS ACROSS 500 REPLICATIONS. THE COVERAGE
LENGTH IS NORMALIZED BY THE MAGNITUDE OF THE CORRESPONDING POINT ESTIMATE.

Case study	$\widehat{v}_0^{\mathrm{hz}}$			$\widehat{v}_0^{\mathrm{vt}}$			$\widehat{v}_0^{\mathrm{mix}}$		
	μ_0^{hz}	μ_0^{vt}	μ_0^{mix}	μ_0^{hz}	μ_0^{vt}	μ_0^{mix}	μ_0^{hz}	μ_0^{vt}	μ_0^{mix}
Basque (CP)	0.92	0.74	0.63	0.99	0.93	0.88	1.00	0.97	0.94
Basque (AL)	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.04
California (CP)	0.95	1.00	0.92	0.64	0.93	0.60	0.98	1.00	0.95
California (AL)	0.07	0.07	0.07	0.03	0.03	0.03	0.08	0.08	0.08
W. Germany (CP)	0.94	1.00	0.93	0.49	0.94	0.49	0.96	1.00	0.95
W. Germany (AL)	0.03	0.03	0.03	0.01	0.01	0.01	0.03	0.03	0.03

5.2.2. *Simulation Results*

For the purposes of stability, we conduct 500 replications of the above DGP for each
study. In the ℓth simulation repeat, we learn the regression coefficients as

$$\widehat{\boldsymbol{\alpha}}^{(\ell)} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \|\boldsymbol{y}_T^{(\ell)} - \boldsymbol{Y}_0 \boldsymbol{\alpha}\|_2^2 \quad \text{and} \quad \widehat{\boldsymbol{\beta}}^{(\ell)} = \operatorname*{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y}_N^{(\ell)} - \boldsymbol{Y}_0' \boldsymbol{\beta}\|_2^2.$$

The corresponding point estimate is defined as $\widehat{Y}_{NT}^{(\ell)}(0) = \langle \boldsymbol{y}_N^{(\ell)}, \widehat{\boldsymbol{\alpha}}^{(\ell)} \rangle = \langle \boldsymbol{y}_T^{(\ell)}, \widehat{\boldsymbol{\beta}}^{(\ell)} \rangle$. We
then construct separate HZ, VT, and mixed homoskedastic confidence intervals based
on (16) and (17) around the point estimate.

In Table II, we report the coverage probabilities (CP) and average lengths (AL)
for each confidence interval with respect to each estimand at the 95% nominal mark.
With respect to μ_0^{hz}, the coverage of the HZ confidence interval is closer to the nominal
coverage than that of the VT and mixed intervals as the latter two can substantially
under- or over-cover. This storyline is consistent for the VT interval with respect to μ_0^{vt}
and the mixed interval with respect to μ_0^{mix}.

Collectively, our formal results and simulations demonstrate that (i) the choice of
estimand directly affects the accuracy of the inference; and (ii) the variance formulas
developed for one estimand may not have the correct coverage for another estimand.
Accordingly, researchers should carefully consider the source of randomness in their
data as it can have a significant influence over their ability to conduct valid inference.
We comment that these conclusions are in line with those drawn in Sekhon and Shem-
Tov (2021), which analyzes the classical difference-in-means estimator with respect to
standard estimands for randomized control trials.

5.3. *Empirical Applications*

Next, we analyze our three case studies of interest. All regression models are built on
pretreatment data only, and the point and variance estimation formulas are separately
applied for the treated unit at each post-treatment period $t > T_0$.

5.3.1. *Point Estimation*

Figure 4 visualizes the counterfactual trajectories generated by the estimators in
Section 3.1. Our findings reinforce Theorems 1 and 2. On a separate note, we observe
that, within the Basque study, the OLS estimates are wildly different from the other
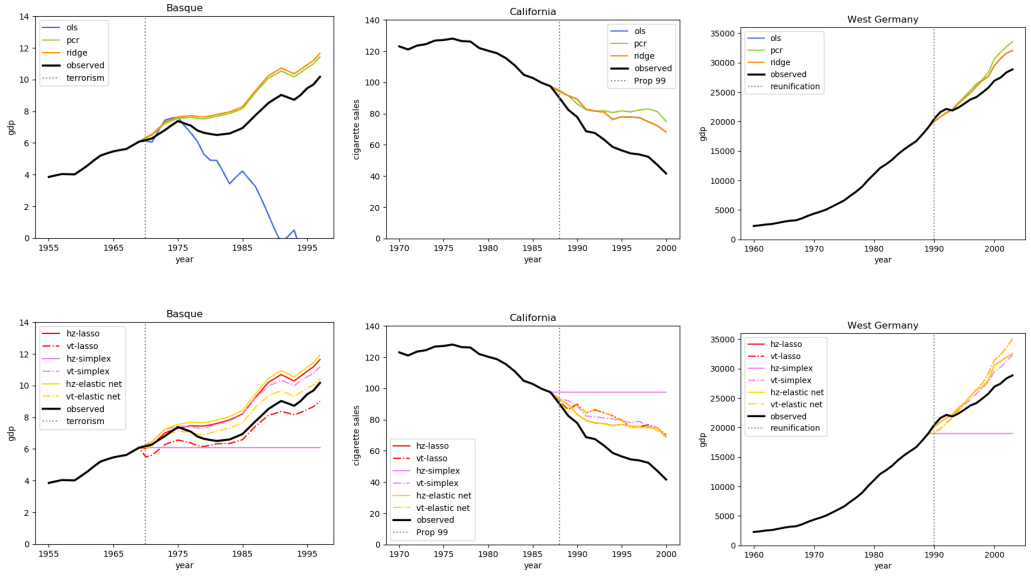
FIGURE 4.—Top and bottom figures correspond to symmetric and asymmetric estimators, respectively. From left to right, the figures are indexed by the Basque, California, and West Germany studies. Across all figures, the treated year is the dotted vertical line; the observed trajectory is in solid black; and the HZ and VT counterfactual trajectories are colored solid and dashed-dotted lines, respectively.

estimates and HZ simplex regression reduces to the last observation carried forward (LOCF) estimator. In the California and West Germany studies, the estimates are all qualitatively similar with the exception of the HZ simplex regression, which again reduces to LOCF. In fact, the OLS and ridge estimates appear to overlap, as well as the lasso and elastic net estimates.

5.3.2. *Inference*

To reduce visual redudancy, we only present the figures associated with the jackknife-based intervals for OLS and PCR in Figures 5 and 6, respectively. Consider the Basque study. The top row of plots in both figures demonstrates that μ_0^{vt} is more accurately estimated than both μ_0^{hz} and μ_0^{mix}. Put differently, there is less uncertainty about conducting inference on μ_0^{vt} relative to the other estimands. At the same time, these plots indicate that if μ_0^{hz} or μ_0^{mix} are the estimands of interest, then the VT confidence interval will undercover in both settings. Analogous statements can be made for the remaining subfigures. As with our simulations, the large potential differences in coverage reinforce the importance of properly reasoning through the source of randomness in the data.

FIGURE 5.—OLS estimates with jackknife confidence intervals. From top to bottom, the rows are indexed by the Basque, California, and West Germany studies. From left to right, the columns are indexed by the HZ, VT, and mixed models. Figure 5 visualizes the problem of degeneracy for OLS-based intervals that is discussed in Remark 7.

6. CONCLUSION

This article rewrites the conventional wisdom on HZ and VT regressions for panel data analysis. Contrary to standard notions, we show that the two regressions yield identical point estimates under several standard settings. At the same time, we articulate that the source of randomness directly affects the accuracy of the inference that can be conducted. From a practical standpoint, this stresses that researchers should carefully consider where the randomness in their data stems from as this decision will then guide their choice of estimand and inferential procedure.

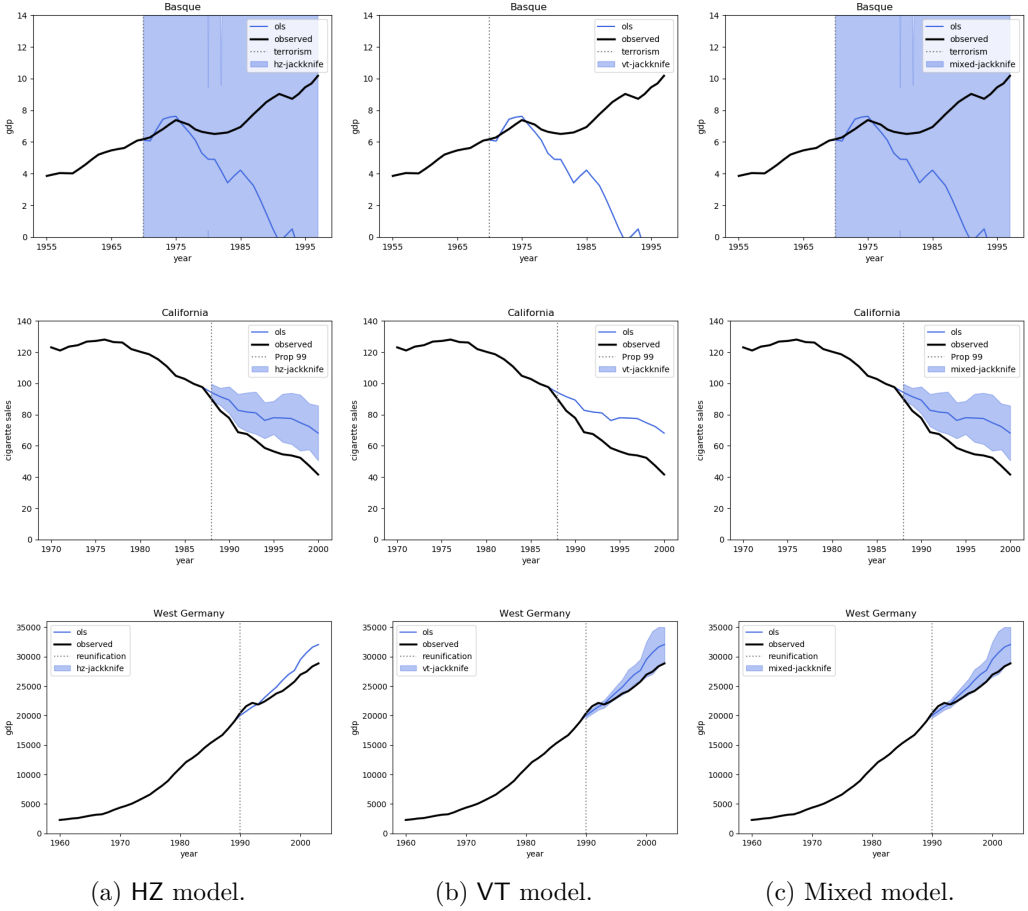(a) HZ model.   (b) VT model.   (c) Mixed model.
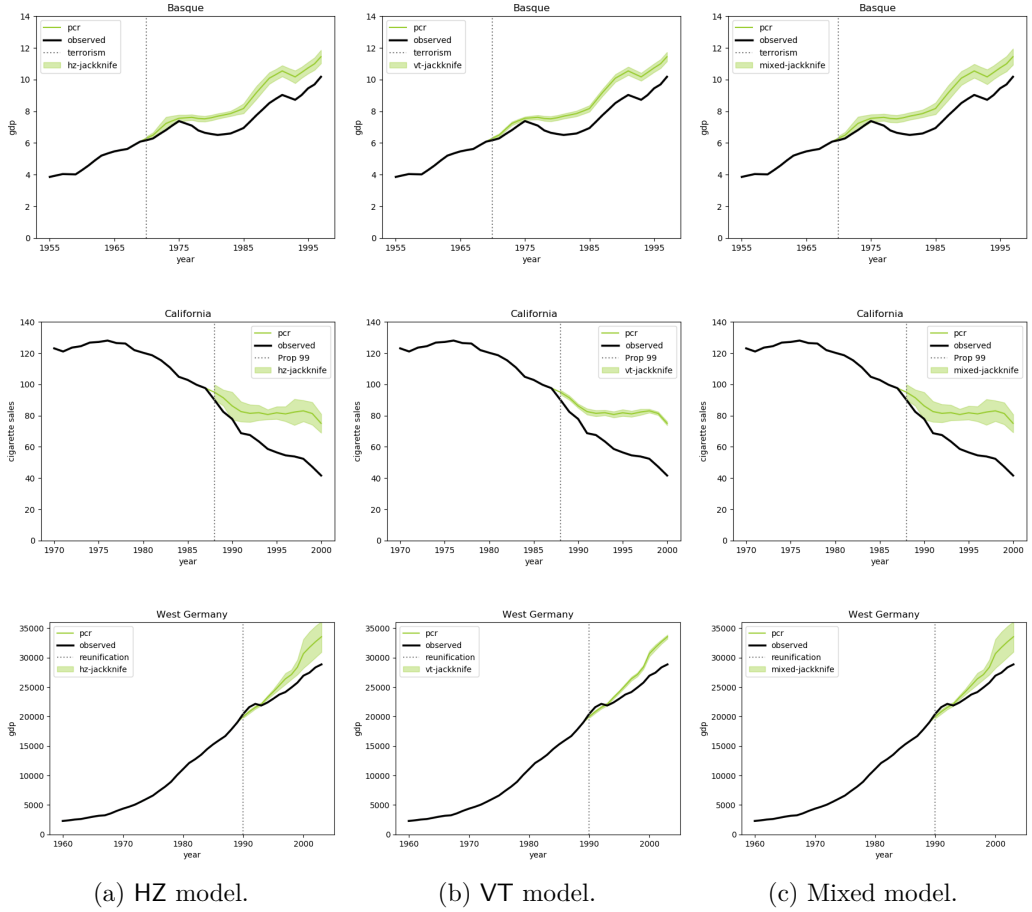
FIGURE 6.—PCR estimates with jackknife confidence intervals. From top to bottom, the rows are indexed by the Basque, California, and West Germany studies. From left to right, the columns are indexed by the HZ, VT, and mixed models.

REFERENCES

ABADIE, ALBERTO (2021): "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59, 391–425. [7]

ABADIE, ALBERTO, SUSAN ATHEY, GUIDO W. IMBENS, AND JEFFREY M. WOOLDRIDGE (2020): "Sampling-Based versus Design-Based Uncertainty in Regression Analysis," *Econometrica*, 88, 265–296. [17]

ABADIE, ALBERTO, ALEXIS DIAMOND, AND JENS HAINMUELLER (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of Californiaâs Tobacco Control Program," *Journal of the American Statistical Association*, 105. [7, 18, 19]

——— (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59, 495–510. [5, 7, 8, 18, 19]

ABADIE, A. AND J. GARDEAZABAL (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93, 113–132. [1, 2, 7]

ABADIE, ALBERTO AND JÉRÉMY L'HOUR (2021): "A Penalized Synthetic Control Estimator for Disaggregated Data," *Journal of the American Statistical Association*, 116, 1817–1834. [7, 9]

AGARWAL, ANISH, DEVAVRAT SHAH, AND DENNIS SHEN (2021): "Synthetic Interventions," *arXiv preprint arXiv:2006.07691.* [6, 27]

Amjad, Muhammad, Vishal Misra, Devavrat Shah, and Dennis Shen (2019): "MRSC: Multi-Dimensional Robust Synthetic Control," *Proc. ACM Meas. Anal. Comput. Syst.*, 3. [6]

Amjad, Muhammad, Devavrat Shah, and Dennis Shen (2018): "Robust Synthetic Control," *Journal of Machine Learning Research*, 19, 1–51. [6]

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager (2021): "Synthetic Difference-in-Differences," *American Economic Review*, 111, 4088–4118. [9]

Ashenfelter, Orley (1978): "Estimating the Effect of Training Programs on Earnings," *The Review of Economics and Statistics*, 60, 47–57. [9]

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi (2021): "Matrix Completion Methods for Causal Panel Data Models," *Journal of the American Statistical Association*, 116, 1716–1730. [2, 5]

Athey, Susan and Guido W. Imbens (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32. [2]

Ben-Michael, Eli, Avi Feller, and Jesse Rothstein (2021): "The Augmented Synthetic Control Method," *Journal of the American Statistical Association*, 116, 1789–1803. [6, 8, 9]

Bottmer, Lea, Guido Imbens, Jann Spiess, and Merrill Warnick (2021): "A Design-Based Perspective on Synthetic Control Methods," . [11, 16, 17, 18]

Carvalho, Carlos, Ricardo Masini, and Marcelo C. Medeiros (2018): "ArCo: An artificial counterfactual approach for high-dimensional panel time-series data," *Journal of Econometrics*, 207, 352–380. [7]

Cattaneo, Matias D., Yingjie Feng, and Rocio Titiunik (2021): "Prediction Intervals for Synthetic Control Methods," *Journal of the American Statistical Association*, 116, 1865–1880. [11, 17]

Chernozhukov, Victor, Kaspar Wuthrich, and Yinchu Zhu (2022): "A *t*-test for synthetic controls," *arXiv preprint arXiv:2006.07691*. [11, 17]

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021): "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association*, 116, 1849–1864. [7, 11, 17]

Cline, Randall E. (1965): "Representations for the Generalized Inverse of Sums of Matrices," *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2, 99–114. [40]

Doudchenko, Nikolay and Guido W Imbens (2016): "Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis," Working Paper 22791, National Bureau of Economic Research. [5, 7]

Gunasekar, Suriya, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro (2017): "Implicit Regularization in Matrix Factorization," in *Advances in Neural Information Processing Systems*. [8]

Hartley, H. O., J. N. K. Rao, and Grace Kiefer (1969): "Variance Estimation with One Unit per Stratum," *Journal of the American Statistical Association*, 64, 841–851. [15, 42]

Hoff, Peter D. (2017): "Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization," *Computational Statistics & Data Analysis*, 115, 186–198. [32, 36]

Hsiao, Cheng, H. Steve Ching, and Shui Ki Wan (2012): "A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China," *Journal of Applied Econometrics*, 27, 705–740. [6]

Imbens, Guido W. and Jeffrey M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [2]

Lehmann, E.L. (2000): "Elements of Large-Sample Theory," . [27, 37]

Li, Kathleen T. (2020): "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*, 115, 2068–2083. [6, 11, 17]

Li, Kathleen T. and David R. Bell (2017): "Estimation of average treatment effects with panel data: Asymptotic theory and implementation," *Journal of Econometrics*, 197, 65–75. [5, 6, 7]

Lin, Winston (2013): "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *The Annals of Applied Statistics*, 7, 295 – 318. [18]

Meyer, Carl D. (1973): "Generalized Inversion of Modified Matrices," *SIAM Journal on Applied Mathematics*, 24, 315–323. [40]

Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro (2015): "In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning," in *International Conference on Learning Representations*. [8]

Rosenbaum, Paul and Donald Rubin (1983): "The Central Role of the Propensity Score in Observational Studies For Causal Effects," *Biometrika*, 70, 41–55. [2]

SEKHON, JASJEET S. AND YOTAM SHEM-TOV (2021): "Inference on a New Class of Sample Average Treatment Effects," *Journal of the American Statistical Association*, 116, 798–804. [17, 20]

SHAO, JUN AND XINWEI DENG (2012): "Estimation in high-dimensional linear models with deterministic design matrices," *The Annals of Statistics*, 40, 812 – 831. [8]

STYAN, GEORGE P.H. (1973): "Hadamard products and multivariate statistical analysis," *Linear Algebra and its Applications*, 6, 217–240. [36]

TIBSHIRANI, RYAN J. (2013): "The lasso problem and uniqueness," *Electronic Journal of Statistics*, 7, 1456 – 1490. [6]

VARGA, RICHARD S. (1962): *Matrix Iterative Analysis*, Prentice-Hall Series in Automatic Computation, Englewood Cliffs: Prentice-Hall. [28]

YU, BIN AND KARL KUMBIER (2020): "Veridical data science," *Proceedings of the National Academy of Sciences*, 117, 3920–3929. [18]

APPENDIX A: Inference

A.1. Model-Based Inference: Asymptotic Properties

First, we state the precise form of Theorem 3. Towards this, let $(\sigma_{iT}^{\mathrm{hz}})^2 = \mathrm{Var}(\varepsilon_{iT}|\boldsymbol{y}_N, \boldsymbol{Y}_0)$ for $i = 1, \ldots, N_0$. We define $((\sigma_{Nt}^{\mathrm{vt}})^2, (\sigma_{iT}^{\mathrm{mix}})^2, (\sigma_{Nt}^{\mathrm{mix}})^2)$ for $i = 1, \ldots, N_0$ and $t = 1, \ldots, T_0$ with respect to $(\boldsymbol{\Sigma}_N^{\mathrm{vt}}, \boldsymbol{\Sigma}_T^{\mathrm{mix}}, \boldsymbol{\Sigma}_N^{\mathrm{mix}})$ analogously.

THEOREM 4: *(i)* [HZ model] *Let Assumptions 1–2 hold. If*

$$\left(\sum_{i \leq N_0} \mathbb{E}\left[|\widehat{\beta}_i \varepsilon_{iT}|^3 \big| \, \boldsymbol{y}_N, \boldsymbol{Y}_0 \right] \right)^2 = o\left(\left(\sum_{i \leq N_0} \widehat{\beta}_i^2 (\sigma_{iT}^{\mathrm{hz}})^2 \right)^3 \right), \tag{21}$$

then

$$\frac{\widehat{Y}_{NT}(0) - \mu_0^{\mathrm{hz}}}{\sqrt{v_0^{\mathrm{hz}}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mu_0^{\mathrm{hz}} = \langle \boldsymbol{y}_N, \boldsymbol{H}^v \boldsymbol{\alpha}^* \rangle$ and $v_0^{\mathrm{hz}} = \widehat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_T^{\mathrm{hz}} \widehat{\boldsymbol{\beta}}$. (ii) [VT model] *Let Assumptions 3–4 hold. If*

$$\left(\sum_{t \leq T_0} \mathbb{E}\left[|\widehat{\alpha}_t \varepsilon_{Nt}|^3 \big| \, \boldsymbol{y}_T, \boldsymbol{Y}_0 \right] \right)^2 = o\left(\left(\sum_{t \leq T_0} \widehat{\alpha}_t^2 (\sigma_{Nt}^{\mathrm{vt}})^2 \right)^3 \right), \tag{22}$$

then

$$\frac{\widehat{Y}_{NT}(0) - \mu_0^{\mathrm{vt}}}{\sqrt{v_0^{\mathrm{vt}}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mu_0^{\mathrm{vt}} = \langle \boldsymbol{y}_T, \boldsymbol{H}^u \boldsymbol{\beta}^* \rangle$ and $v_0^{\mathrm{vt}} = \widehat{\boldsymbol{\alpha}}' \boldsymbol{\Sigma}_N^{\mathrm{vt}} \widehat{\boldsymbol{\alpha}}$. (iii) [Mixed model] *Let Assumptions 5–6 hold. If*

$$\left(\sum_{i \leq N_0} \sum_{t \leq T_0} \mathbb{E}\left[|(\boldsymbol{Y}_0^\dagger)_{it} \left\{ \mathbb{E}[Y_{iT}|\boldsymbol{Y}_0]\varepsilon_{Nt} + \mathbb{E}[Y_{Nt}|\boldsymbol{Y}_0]\varepsilon_{iT} + \varepsilon_{iT}\varepsilon_{Nt} \right\}|^3 \, \big| \, \boldsymbol{Y}_0 \right] \right)^2$$
$$= o\left(\left(\sum_{i \leq N_0} \sum_{t \leq T_0} (\boldsymbol{Y}_0^\dagger)_{it}^2 \left\{ \mathbb{E}[Y_{iT}|\boldsymbol{Y}_0]^2 (\sigma_{Nt}^{\mathrm{mix}})^2 + \mathbb{E}[Y_{Nt}|\boldsymbol{Y}_0]^2 (\sigma_{iT}^{\mathrm{mix}})^2 + (\sigma_{iT}^{\mathrm{mix}})^2 (\sigma_{Nt}^{\mathrm{mix}})^2 \right\} \right)^3 \right), \tag{23}$$

then

$$\frac{\widehat{Y}_{NT}(0) - \mu_0^{\mathrm{mix}}}{\sqrt{v_0^{\mathrm{mix}}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mu_0^{\mathrm{mix}} = \langle \boldsymbol{\alpha}^*, \boldsymbol{Y}_0' \boldsymbol{\beta}^* \rangle$ and

$$v_0^{\mathrm{mix}} = (\boldsymbol{H}^u \boldsymbol{\beta}^*)' \boldsymbol{\Sigma}_T^{\mathrm{mix}} (\boldsymbol{H}^u \boldsymbol{\beta}^*) + (\boldsymbol{H}^v \boldsymbol{\alpha}^*)' \boldsymbol{\Sigma}_N^{\mathrm{mix}} (\boldsymbol{H}^v \boldsymbol{\alpha}^*) + \mathrm{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\mathrm{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\mathrm{mix}}).$$

If $(\varepsilon_{iT}, (\sigma_{iT}^{\mathrm{hz}})^2)$ are bounded, then (21) translates to $\sum_{i \leq N_0} |\widehat{\beta}_i|^3 = o(\|\widehat{\boldsymbol{\beta}}\|_2^3)$, which rules out outlier coefficients; a similar interpretation can be derived for (22). Similarly, if $(\varepsilon_{iT}, \varepsilon_{Nt})$ and $(\sigma_{iT}^2, \sigma_{Nt}^2)$ are bounded for all (i,t), then (23) loosely translates to

$$\sum_{i \leq N_0} |\widehat{\beta}_i|^3 + \sum_{t \leq T_0} |\widehat{\alpha}_t|^3 + \sum_{i \leq N_0} \sum_{t \leq T_0} |(\boldsymbol{Y}_0^\dagger)_{it}|^3 = o\left(\|\widehat{\boldsymbol{\beta}}\|_2^3 + \|\widehat{\boldsymbol{\alpha}}\|_2^3 + \|\boldsymbol{Y}_0^\dagger\|_F^3\right),$$

which effectively bounds the magnitudes of the HZ and VT OLS coefficients and pseudoinverse matrix entries. We note that (21)–(23) are known as Lyapunov's condition, and we refer the interested reader to Lehmann (2000) for details.

Next, we provide a bound on the trace term in v_0^{mix}. Beginning with the upper bound, notice that

$$\mathrm{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\mathrm{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\mathrm{mix}}) \leq \max_{i \leq N_0}(\sigma_{iT}^{\mathrm{mix}})^2 \max_{t \leq T_0}(\sigma_{Nt}^{\mathrm{mix}})^2 \, \mathrm{tr}(\boldsymbol{Y}_0^\dagger (\boldsymbol{Y}_0')^\dagger).$$

By the cyclic property of the trace operator,

$$\mathrm{tr}(\boldsymbol{Y}_0^\dagger (\boldsymbol{Y}_0')^\dagger) = \mathrm{tr}(\boldsymbol{V} \boldsymbol{S}^{-2} \boldsymbol{V}') = \mathrm{tr}(\boldsymbol{S}^{-2} \boldsymbol{V}' \boldsymbol{V}) = \mathrm{tr}(\boldsymbol{S}^{-2}).$$

Putting everything together, we obtain

$$\mathrm{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\mathrm{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\mathrm{mix}}) \leq \max_{i \leq N_0}(\sigma_{iT}^{\mathrm{mix}})^2 \max_{t \leq T_0}(\sigma_{Nt}^{\mathrm{mix}})^2 \, \mathrm{tr}(\boldsymbol{S}^{-2}).$$

The same arguments can be applied to derive the lower bound, which yields

$$\mathrm{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\mathrm{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\mathrm{mix}})$$
$$\in \mathrm{tr}(\boldsymbol{S}^{-2}) \left[\min_{i \leq N_0}(\sigma_{iT}^{\mathrm{mix}})^2 \min_{t \leq T_0}(\sigma_{Nt}^{\mathrm{mix}})^2, \; \max_{i \leq N_0}(\sigma_{iT}^{\mathrm{mix}})^2 \max_{t \leq T_0}(\sigma_{Nt}^{\mathrm{mix}})^2 \right]. \tag{24}$$

From (24), we see that the lower and upper bounds match under homoskedasticity.

A.2. Model-Based Inference: Confidence Intervals

Next, we discuss technical aspects of the variance estimators in Section 4.1.3.

Homoskedastic errors. We take note of the recent work of Agarwal et al. (2021) in the synthetic controls literature. Agarwal et al. (2021) propose a VT PCR estimator under the homoskedastic setting and provide a similar confidence interval to that of (17) via large-sample approximations. Under a closely related VT model, they propose $\widehat{\boldsymbol{\beta}}' \widehat{\boldsymbol{\Sigma}}_N^{\mathrm{homo}} \widehat{\boldsymbol{\beta}}$ in place of $\widehat{\boldsymbol{\alpha}}' \widehat{\boldsymbol{\Sigma}}_N^{\mathrm{homo}} \widehat{\boldsymbol{\alpha}}$. While the point estimate of Agarwal et al. (2021) also takes the form $\langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle$, their variance estimator only depends on $(\boldsymbol{y}_N, \boldsymbol{Y}_0)$; in comparison, ours depends on $(\boldsymbol{y}_N, \boldsymbol{y}_T, \boldsymbol{Y}_0)$. Therefore, the confidence interval as per Agarwal et al. (2021) is numerically identical for every post-treatment point estimate while ours can vary across the post-treatment periods, which may be favorable.

Heteroskedastic errors. Consider the heteroskedastic setting.

I: Jackknife. In the following lemma, we quantify the bias in Lemma 2.

Lemma 4—Detailed restatement of Lemma 2: *(i)* [HZ model] *Let Assumptions 1–2 hold. If $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})$ is nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\text{hz}} + \boldsymbol{\Delta}^{\text{hz}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{hz,jack}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = v_0^{\text{hz}} + \widehat{\boldsymbol{\alpha}}' \boldsymbol{\Delta}^{\text{hz}} \widehat{\boldsymbol{\alpha}},$$

where $\Delta_{\ell\ell}^{\text{hz}} = \sum_{j\neq\ell} (\sigma_{jT}^{\text{hz}})^2 (H_{\ell j}^u)^2 (1-H_{\ell\ell}^u)^{-2}$ for $\ell = 1,\ldots,N_0$. *(ii)* [VT model] *Let Assumptions 3–4 hold. If $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v \circ \boldsymbol{I})$ is nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{jack}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\text{vt}} + \boldsymbol{\Gamma}^{\text{vt}} \quad and \quad \mathbb{E}[\widehat{v}_0^{\text{vt,jack}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = v_0^{\text{vt}} + \widehat{\boldsymbol{\beta}}' \boldsymbol{\Gamma}^{\text{vt}} \widehat{\boldsymbol{\beta}},$$

where $\Gamma_{\ell\ell}^{\text{vt}} = \sum_{j\neq\ell} (\sigma_{Nj}^{\text{vt}})^2 (H_{\ell j}^v)^2 (1-H_{\ell\ell}^v)^{-2}$ for $\ell = 1,\ldots,T_0$. *(iii)* [Mixed model] *Let Assumptions 5–6 hold. If $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})$ and $(\boldsymbol{H}_\perp^v \circ \boldsymbol{H}_\perp^v \circ \boldsymbol{I})$ are nonsingular, then*

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\text{mix}} + \boldsymbol{\Delta}^{\text{mix}}, \quad \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{jack}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\text{mix}} + \boldsymbol{\Gamma}^{\text{mix}},$$

$$\mathbb{E}[\widehat{v}^{\text{mix,jack}}(\boldsymbol{Y}_0)|\boldsymbol{Y}_0]$$
$$= v_0^{\text{mix}} + (\boldsymbol{H}^u \boldsymbol{\beta}^*)' \boldsymbol{\Delta}^{\text{mix}} (\boldsymbol{H}^u \boldsymbol{\beta}^*) + (\boldsymbol{H}^v \boldsymbol{\alpha}^*)' \boldsymbol{\Gamma}^{\text{mix}} (\boldsymbol{H}^v \boldsymbol{\alpha}^*) + \text{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Delta}^{\text{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Gamma}^{\text{mix}}),$$

where $\Delta_{\ell\ell}^{\text{mix}}$ and $\Gamma_{\ell\ell}^{\text{mix}}$ are defined analogously to $\Delta_{\ell\ell}^{\text{hz}}$ and $\Gamma_{\ell\ell}^{\text{vt}}$, respectively, with $(\sigma_{jT}^{\text{mix}})^2$ and $(\sigma_{Nj}^{\text{mix}})^2$ in place of $(\sigma_{jT}^{\text{hz}})^2$ and $(\sigma_{Nj}^{\text{vt}})^2$, respectively.

Without loss of generality, we consider the magnitudes of $(\Delta_{\ell\ell}^{\text{mix}}, \Gamma_{\ell\ell}^{\text{mix}})$. Towards bounding the former quantity, notice that $\boldsymbol{H}^u$ is an orthogonal projector and is thus idempotent, i.e., $(\boldsymbol{H}^u)^2 = \boldsymbol{H}^u$, and symmetric. Therefore,

$$H_{\ell\ell}^u = (H_{\ell\ell}^u)^2 + \sum_{j\neq\ell}(H_{\ell j}^u)^2 \implies \sum_{j\neq\ell}(H_{\ell j}^u)^2 = H_{\ell\ell}^u(1-H_{\ell\ell}^u). \tag{25}$$

This yields $\Delta_{\ell\ell}^{\text{mix}} \in H_{\ell\ell}^u(1-H_{\ell\ell}^u)^{-1}[\min_{j\neq\ell}(\sigma_{jT}^{\text{mix}})^2, \max_{j\neq\ell}(\sigma_{jT}^{\text{mix}})^2]$. Since $\sum_{j\neq\ell}(H_{\ell j}^u)^2 \geq 0$, (25) implies $H_{\ell\ell}^u \in [0,1]$. Thus, $\Delta_{\ell\ell}^{\text{mix}} = 0$ if $H_{\ell\ell}^u = 0$ and diverges if $H_{\ell\ell}^u = 1$. If $H_{\ell\ell}^u$ takes the average value $N_0^{-1}\sum_\ell H_{\ell\ell}^u = RN_0^{-1}$, then it follows that $\Delta_{\ell\ell}^{\text{mix}} \in R(N_0 - R)^{-1}[\min_{j\neq\ell}(\sigma_{jT}^{\text{mix}})^2, \max_{j\neq\ell}(\sigma_{jT}^{\text{mix}})^2]$. A similar result is derived for $\Gamma_{\ell\ell}^{\text{mix}}$.

II: HRK-estimator. Recall Lemma 3. Consider the HZ estimator and the invertibility of $(\boldsymbol{H}^u \circ \boldsymbol{H}^u)$. A sufficient condition is strict diagonal dominance (Varga, 1962): $(1 - H_{\ell\ell}^u)^2 > \sum_{j\neq\ell}(H_{\ell j}^u)^2$. Using (25), we simplify this condition as $(1-H_{\ell\ell}^u)^2 > H_{\ell\ell}^u - (H_{\ell\ell}^u)^2$. Thus, $\max_\ell H_{\ell\ell}^u < 1/2$ is a sufficient condition for invertibility. The same arguments apply for VT regression.

Mixed variance estimator. Let us bound $\widehat{v}_0^{\text{mix}}$. From (24), it follows that $\widehat{v}_0^{\text{mix}} \in [\widehat{v}_{0,\min}^{\text{mix}}, \widehat{v}_{0,\max}^{\text{mix}}]$, where

$$\widehat{v}_{0,\min}^{\text{mix}} = \min_{i\leq N_0} \widehat{\sigma}_{iT}^2 \|\widehat{\boldsymbol{\beta}}\|_2^2 + \min_{t\leq T_0} \widehat{\sigma}_{Nt}^2 \|\widehat{\boldsymbol{\alpha}}\|_2^2 - \max_{i\leq N_0}\widehat{\sigma}_{iT}^2 \max_{t\leq T_0}\widehat{\sigma}_{Nt}^2 \text{tr}(\boldsymbol{S}^{-2}),$$

$$\widehat{v}_{0,\max}^{\text{mix}} = \max_{i\leq N_0} \widehat{\sigma}_{iT}^2 \|\widehat{\boldsymbol{\beta}}\|_2^2 + \max_{t\leq T_0} \widehat{\sigma}_{Nt}^2 \|\widehat{\boldsymbol{\alpha}}\|_2^2 - \min_{i\leq N_0}\widehat{\sigma}_{iT}^2 \min_{t\leq T_0}\widehat{\sigma}_{Nt}^2 \text{tr}(\boldsymbol{S}^{-2}).$$

Here, $\widehat{\sigma}_{iT}^2$ and $\widehat{\sigma}_{Nt}^2$ are the ith and tth diagonal elements of $\widehat{\boldsymbol{\Sigma}}_T$ and $\widehat{\boldsymbol{\Sigma}}_N$, respectively. Observe that $\widehat{v}_{0,\min}^{\text{mix}} = \widehat{v}_{0,\max}^{\text{mix}}$ for homoskedastic errors.

APPENDIX B: Illustrations

This section provides details on the simulations that were absent in the main article.

B.1. Implementation Details

For ridge, lasso, and elastic net regressions, we use the default `scikit-learn` hyperparameters (λ_1, λ_2). For PCR, we choose the number of principal components k via the approach described in Section 5.2. This yields $k =$ for the Basque study, $k = 3$ for the California study, and $k = 4$ for the West Germany study. We implement simplex regression using the code made available at https://matheusfacure.github.io/python-causality-handbook/15-Synthetic-Control.html.

B.2. Data-Inspired Simulation Studies

Formally, we define average length (AL) as

$$\text{AL} = \frac{1}{m} \sum_{\ell=1}^{m} \frac{(2 \cdot 1.96)\sqrt{\widehat{v}_0^{(\ell)}}}{|\widehat{Y}_{NT}^{(\ell)}(0)|},$$

where $\widehat{v}_0^{(\ell)}$ and $\widehat{Y}_{NT}^{(\ell)}(0)$ are the variance and point estimates for the ℓth repeat.

APPENDIX C: Proofs for Point Estimation

Helper lemmas. To establish Theorems 1 and 2, we first state the following useful lemmas for the collection of regression formulations presented in Section 3. We provide their proofs in Appendix C.6.

LEMMA 5—OLS: $\mathsf{HZ} = \mathsf{VT}$ *for OLS with* $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ *as the minimum ℓ_2-norm solutions:*

$$\widehat{Y}_{NT}^{\text{hz}}(0) = \widehat{Y}_{NT}^{\text{vt}}(0) = \langle \boldsymbol{y}_N, \boldsymbol{Y}_0^{\dagger} \boldsymbol{y}_T \rangle = \sum_{\ell=1}^{R} (1/s_\ell) \langle \boldsymbol{y}_N, \boldsymbol{v}_\ell \rangle \langle \boldsymbol{u}_\ell, \boldsymbol{y}_T \rangle.$$

LEMMA 6—PCR: $\mathsf{HZ} = \mathsf{VT}$ *for PCR with the same choice of $k < R$:*

$$\widehat{Y}_{NT}^{\text{hz}}(0) = \widehat{Y}_{NT}^{\text{vt}}(0) = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0^{(k)})^{\dagger} \boldsymbol{y}_T \rangle = \sum_{\ell=1}^{k} (1/s_\ell) \langle \boldsymbol{y}_N, \boldsymbol{v}_\ell \rangle \langle \boldsymbol{u}_\ell, \boldsymbol{y}_T \rangle.$$

LEMMA 7—Ridge: $\mathsf{HZ} = \mathsf{VT}$ *for ridge regression with the same choice of $\lambda_2 > 0$:*

$$\widehat{Y}_{NT}^{\text{hz}}(0) = \widehat{Y}_{NT}^{\text{vt}}(0) = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0' \boldsymbol{Y}_0 + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{Y}_0' \boldsymbol{y}_T \rangle = \sum_{\ell=1}^{R} \frac{s_\ell}{s_\ell^2 + \lambda_2} \langle \boldsymbol{y}_N, \boldsymbol{v}_\ell \rangle \langle \boldsymbol{u}_\ell, \boldsymbol{y}_T \rangle.$$

LEMMA 8—Lasso: $\mathsf{HZ} \neq \mathsf{VT}$ *for lasso regression.*

LEMMA 9—Elastic net: $\mathsf{HZ} \neq \mathsf{VT}$ *for elastic net regression.*

LEMMA 10—Simplex regression: $\mathsf{HZ} \neq \mathsf{VT}$ *for simplex regression.*

C.1. Proof of Theorem 1

PROOF: The proof is immediate from Lemmas 5–7.   *Q.E.D.*

C.2. Proof of Theorem 2

PROOF: The proof is immediate from Lemmas 8–10.   *Q.E.D.*

C.3. Proof of Corollary 1

PROOF: We begin with the OLS. Recall that $\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle$ with $\widehat{\boldsymbol{\alpha}} = \boldsymbol{Y}_0^\dagger \boldsymbol{y}_T$ and $\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle$ with $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}_0')^\dagger \boldsymbol{y}_N$. By Theorem 1, we have

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_N, \boldsymbol{Y}_0^\dagger \boldsymbol{y}_T \rangle.$$

Returning to (6), we obtain

$$\widehat{Y}_{NT}^{\mathrm{sdid}}(0) = \widehat{Y}_{NT}^{\mathrm{asc}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle + \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle - \langle \widehat{\boldsymbol{\alpha}}, \boldsymbol{Y}_0' \widehat{\boldsymbol{\beta}} \rangle = 2 \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle - \langle \widehat{\boldsymbol{\alpha}}, \boldsymbol{Y}_0' \widehat{\boldsymbol{\beta}} \rangle. \quad (26)$$

Recall $(\boldsymbol{Y}_0')^\dagger = (\boldsymbol{Y}_0^\dagger)'$. Therefore,

$$\langle \widehat{\boldsymbol{\alpha}}, \boldsymbol{Y}_0' \widehat{\boldsymbol{\beta}} \rangle = \boldsymbol{y}_T' (\boldsymbol{Y}_0')^\dagger \boldsymbol{Y}_0' (\boldsymbol{Y}_0')^\dagger \boldsymbol{y}_N = \boldsymbol{y}_T' (\boldsymbol{Y}_0')^\dagger \boldsymbol{y}_N = \boldsymbol{y}_T' \widehat{\boldsymbol{\beta}}. \quad (27)$$

Plugging (27) into (26), we conclude

$$\widehat{Y}_{NT}^{\mathrm{sdid}}(0) = \widehat{Y}_{NT}^{\mathrm{asc}}(0) = 2 \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle - \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle = \widehat{Y}_{NT}^{\mathrm{vt}}(0) = \widehat{Y}_{NT}^{\mathrm{hz}}(0).$$

Now, observe that the same arguments above hold when $\boldsymbol{Y}_0^{(k)}$ takes the place of $\boldsymbol{Y}_0$ for any $k < R$. Therefore, the same reduction can be derived for PCR.   *Q.E.D.*

C.4. Proof of Corollary 2

PROOF: Let $\boldsymbol{Y}_0^{\mathrm{hz}} = [\boldsymbol{1}, \boldsymbol{Y}_0]$ and $\boldsymbol{Y}_0^{\mathrm{vt}} = [\boldsymbol{1}, \boldsymbol{Y}_0']$. The proof is immediate from Theorem 1 by noting that $(\boldsymbol{Y}_0^{\mathrm{hz}})' \neq \boldsymbol{Y}_0^{\mathrm{vt}}$.   *Q.E.D.*

C.5. Proof of Corollary 3

PROOF: Consider HZ ridge regression. To begin, the optimality conditions give

$$\nabla_{(\alpha_0, \alpha_1, \boldsymbol{\alpha})} \left\{ \|\boldsymbol{y}_T - \boldsymbol{Y}_0 \boldsymbol{\alpha} - \alpha_0 \boldsymbol{1}\|_2^2 + \|\boldsymbol{y}_N - \boldsymbol{\alpha}_1 \boldsymbol{1}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_2^2 \right\} = 0.$$

Solving for α_0, we have $\widehat{\alpha}_0 = (1/N_0) \langle \boldsymbol{y}_T, \boldsymbol{1} \rangle$, where we have used the fact that $\boldsymbol{Y}_0' \boldsymbol{1} = \boldsymbol{0}$. Solving for α_1, we have $\widehat{\alpha}_1 = (1/T_0) \langle \boldsymbol{y}_N, \boldsymbol{1} \rangle$. Finally, solving for $\boldsymbol{\alpha}$, we have $\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Y}_0' \boldsymbol{Y}_0 + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}_0' \boldsymbol{y}_T$.

Switching gears to VT ridge regression, similar arguments yield $\widehat{\beta}_0 = (1/T_0) \langle \boldsymbol{y}_N, \boldsymbol{1} \rangle$, $\widehat{\beta}_1 = (1/N_0) \langle \boldsymbol{y}_T, \boldsymbol{1} \rangle$, and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda \boldsymbol{I}) \boldsymbol{Y}_0 \boldsymbol{y}_N$.

We establish our desired result by invoking Theorem 1 to obtain $\langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle$. Next, we observe that $\widehat{\alpha}_0 = \widehat{\beta}_1$ and $\widehat{\beta}_0 = \widehat{\alpha}_1$. This proves that HZ and VT ridge yield numerically identical point estimates. Setting $\lambda = 0$ and using the pseudoinverse, we have our result for OLS. The result for PCR is then established from the OLS result by substituting $\boldsymbol{Y}_0^{(k)}$ for $\boldsymbol{Y}_0$. This completes the proof.   *Q.E.D.*

C.6. Proofs of Point Estimation Lemmas

C.6.1. Proof of Lemma 5: OLS

PROOF: Consider HZ regression. By the optimality conditions,

$$\nabla_{\boldsymbol{\alpha}} \| \boldsymbol{y}_T - \boldsymbol{Y}_0 \boldsymbol{\alpha} \|_2^2 = 0.$$

Solving for $\boldsymbol{\alpha}$, we derive the well-known "normal equations"

$$\boldsymbol{Y}_0' \boldsymbol{Y}_0 \boldsymbol{\alpha} = \boldsymbol{Y}_0' \boldsymbol{y}_T.$$

Using the pseudoinverse, we obtain

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Y}_0' \boldsymbol{Y}_0)^{\dagger} \boldsymbol{Y}_0' \boldsymbol{y}_T = \boldsymbol{Y}_0^{\dagger} \boldsymbol{y}_T$$

Observe that this corresponds to the unique minimum ℓ_2-norm solution that lies within the rowspace of $\boldsymbol{Y}_0$. Therefore, the HZ prediction is given by

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle = \langle \boldsymbol{y}_N, \boldsymbol{Y}_0^{\dagger} \boldsymbol{y}_T \rangle.$$

Following the arguments above for VT regression, it follows that

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}_0 \boldsymbol{Y}_0')^{\dagger} \boldsymbol{Y}_0 \boldsymbol{y}_N = (\boldsymbol{Y}_0')^{\dagger} \boldsymbol{y}_N,$$

which corresponds to the unique minimum ℓ_2-norm solution that lies within the columnspace of $\boldsymbol{Y}_0$. Therefore,

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle = \langle \boldsymbol{y}_T, (\boldsymbol{Y}_0')^{\dagger} \boldsymbol{y}_N \rangle.$$

Given that $(\boldsymbol{Y}_0')^{\dagger} = (\boldsymbol{Y}_0^{\dagger})'$, we conclude

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \boldsymbol{Y}_0^{\dagger} \boldsymbol{y}_T \rangle = \langle \boldsymbol{y}_T, (\boldsymbol{Y}_0')^{\dagger} \boldsymbol{y}_N \rangle = \widehat{Y}_{NT}^{\mathrm{vt}}(0).$$

Q.E.D.

C.6.2. Proof of Lemma 6: PCR

PROOF: Consider HZ regression with any $k < R$. Let $\boldsymbol{U}_k \in \mathbb{R}^{N_0 \times k}$ and $\boldsymbol{V}_k \in \mathbb{R}^{T_0 \times k}$ denote the matrices formed by the top k left and right singular vectors, respectively, and $\boldsymbol{S}_k \in \mathbb{R}^{k \times k}$ denote the matrix of top k singular values. Observe that

$$\left((\boldsymbol{Y}_0^{(k)})' \boldsymbol{Y}_0^{(k)} \right)^{\dagger} (\boldsymbol{Y}_0^{(k)})' = (\boldsymbol{V}_k \boldsymbol{S}_k^{-2} \boldsymbol{V}_k') \boldsymbol{V}_k \boldsymbol{S}_k \boldsymbol{U}_k' = \boldsymbol{V}_k \boldsymbol{S}_k^{-1} \boldsymbol{U}_k' = (\boldsymbol{Y}_0^{(k)})^{\dagger}.$$

Therefore, $\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Y}_0^{(k)})^{\dagger} \boldsymbol{y}_T$, which corresponds to the unique minimum ℓ_2-norm solution that lies within the rowspace of $\boldsymbol{Y}_0^{(k)}$. Following the proof of Lemma 5, we conclude that

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0^{(k)})^{\dagger} \boldsymbol{y}_T \rangle.$$

Similarly, for VT regression, we note that

$$\left(\boldsymbol{Y}_0^{(k)} (\boldsymbol{Y}_0^{(k)})' \right)^{\dagger} \boldsymbol{Y}_0^{(k)} = (\boldsymbol{U}_k \boldsymbol{S}_k^{-2} \boldsymbol{U}_k') \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k' = \boldsymbol{U}_k \boldsymbol{S}_k^{-1} \boldsymbol{V}_k' = ((\boldsymbol{Y}_0^{(k)})')^{\dagger}.$$

In turn, we have $\widehat{\boldsymbol{\beta}} = ((\boldsymbol{Y}_0^{(k)})')^{\dagger}\boldsymbol{y}_N$, which corresponds to the unique minimum ℓ_2-norm solution that lies within the columnspace of $\boldsymbol{Y}_0^{(k)}$. Moreover,

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle = \langle \boldsymbol{y}_T, ((\boldsymbol{Y}_0^{(k)})')^{\dagger}\boldsymbol{y}_N \rangle.$$

We finish by establishing

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0^{(k)})^{\dagger}\boldsymbol{y}_T \rangle = \langle \boldsymbol{y}_T, ((\boldsymbol{Y}_0^{(k)})')^{\dagger}\boldsymbol{y}_N \rangle = \widehat{Y}_{NT}^{\mathrm{vt}}(0).$$

$$Q.E.D.$$

Helper lemma. To establish Lemmas 7–9, we first establish a general result in Lemma 11 for ℓ_p-penalties, where $p = 2/K$ and K is an integer ≥ 1, based on the contributions of Hoff (2017). More formally, consider

(a) HZ regression: for $K \geq 1$ and $\lambda > 0$,

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_p^p \tag{28}$$

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle.$$

(b) VT regression: for $K \geq 1$ and $\lambda > 0$,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y}_N - \boldsymbol{Y}_0'\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_p^p \tag{29}$$

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}} \rangle.$$

We remark that $K = 1$ and $K = 2$ yield ridge and lasso regression, respectively, while $K > 2$ yields non-convex penalties. We relegate the proof of Lemma 11 to Appendix C.6.7.

LEMMA 11: *For any $K \geq 1$ and $\lambda > 0$, a HZ and VT regression solution is*

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}}_1 \circ \cdots \circ \widehat{\boldsymbol{\alpha}}_K \rangle$$
$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}}_1 \circ \cdots \circ \widehat{\boldsymbol{\beta}}_K \rangle,$$

where for every $k \leq K$,

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k})\boldsymbol{Y}_0'\boldsymbol{Y}_0\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k}) + \frac{\lambda}{K}\boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k})\boldsymbol{Y}_0'\boldsymbol{y}_T,$$

$$\widehat{\boldsymbol{\beta}}_k = \left(\boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k})\boldsymbol{Y}_0\boldsymbol{Y}_0'\boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k}) + \frac{\lambda}{K}\boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k})\boldsymbol{Y}_0\boldsymbol{y}_N,$$

$\widehat{\boldsymbol{\alpha}}_{\sim k} = \widehat{\boldsymbol{\alpha}}_1 \circ \cdots \circ \widehat{\boldsymbol{\alpha}}_{k-1} \circ \widehat{\boldsymbol{\alpha}}_{k+1} \circ \cdots \circ \widehat{\boldsymbol{\alpha}}_K$, $\widehat{\boldsymbol{\beta}}_{\sim k} = \widehat{\boldsymbol{\beta}}_1 \circ \cdots \circ \widehat{\boldsymbol{\beta}}_{k-1} \circ \widehat{\boldsymbol{\beta}}_{k+1} \circ \cdots \circ \widehat{\boldsymbol{\beta}}_K$, and $\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k})$ and $\boldsymbol{D}(\widehat{\boldsymbol{\beta}} \sim k)$ are diagonal matrices formed from $\widehat{\boldsymbol{\alpha}}_{\sim k}$ and $\widehat{\boldsymbol{\beta}}_{\sim k}$, respectively.

C.6.3. *Proof of Lemma 7: Ridge Regression*

PROOF: By Lemma 11 for $K = 1$ and $\lambda = \lambda_2 > 0$, the HZ regression solution is

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, (\boldsymbol{Y}_0' \boldsymbol{Y}_0 + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{Y}_0' \boldsymbol{y}_T \rangle.$$

Similarly, the VT regression solution is given by

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{Y}_0 \boldsymbol{y}_N \rangle.$$

Since $(\boldsymbol{Y}_0' \boldsymbol{Y}_0 + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}_0' = \boldsymbol{Y}_0' (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda \boldsymbol{I})^{-1}$, it follows that $\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \widehat{Y}_{NT}^{\mathrm{vt}}(0)$. *Q.E.D.*

C.6.4. *Proof of Lemma 8: Lasso Regression*

PROOF: By Lemma 11 for $K = 2$ and $\lambda = \lambda_1 > 0$, a HZ regression solution is

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}}_1 \circ \widehat{\boldsymbol{\alpha}}_2 \rangle, \tag{30}$$

where

$$\widehat{\boldsymbol{\alpha}}_{1+k} = \left(\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{2-k}) \boldsymbol{Y}_0' \boldsymbol{Y}_0 \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{2-k}) + \frac{\lambda_1}{2} \boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{2-k}) \boldsymbol{Y}_0' \boldsymbol{y}_T$$

for $k \in \{0, 1\}$. Similarly, a VT regression solution is given by

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}}_1 \circ \widehat{\boldsymbol{\beta}}_2 \rangle, \tag{31}$$

where

$$\widehat{\boldsymbol{\beta}}_{1+k} = \left(\boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{2-k}) \boldsymbol{Y}_0 \boldsymbol{Y}_0' \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{2-k}) + \frac{\lambda_1}{2} \boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{2-k}) \boldsymbol{Y}_0 \boldsymbol{y}_N$$

for $k \in \{0, 1\}$. Leveraging (30) and (31), we find that the HZ regression solution can be linear in y and at least quadratic in q. On the other hand, the VT regression solution can be linear in q and at least quadratic in y. Since the lasso solution is unique under the assumption the entries of $\boldsymbol{Y}_0$ are drawn from a continuous distribution, this implies that HZ and VT regressions do not yield matching solutions in general.   *Q.E.D.*

C.6.5. *Proof of Lemma 9: Elastic Net Regression*

PROOF: Consider HZ regression. We rewrite (1) in a lasso formulation:

$$\widehat{\boldsymbol{\alpha}}^* = \operatorname*{argmin}_{\boldsymbol{\alpha}^*} \| \boldsymbol{y}_T^* - \boldsymbol{Y}_0^* \boldsymbol{\alpha}^* \|_2^2 + \lambda^* \| \boldsymbol{\alpha}^* \|_1, \tag{32}$$

where

$$q^* = \begin{pmatrix} \boldsymbol{y}_T \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{Y}_0^* = \frac{1}{\sqrt{1 + \lambda_2}} \begin{pmatrix} \boldsymbol{Y}_0 \\ \sqrt{\lambda_2} \boldsymbol{I} \end{pmatrix}, \quad \lambda^* = \frac{\lambda_1}{\sqrt{1 + \lambda_2}}, \quad \boldsymbol{\alpha}^* = (\sqrt{1 + \lambda_2}) \boldsymbol{\alpha}.$$

We apply Lemma 11 to (32) with $K = 2$ and $\lambda = \lambda^* > 0$ to obtain

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \frac{\langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}}_1^* \circ \widehat{\boldsymbol{\alpha}}_2^* \rangle}{\sqrt{1 + \lambda_2}}, \tag{33}$$

where

$$\widehat{\boldsymbol{\alpha}}^*_{1+k} = \left(\frac{1}{\sqrt{1+\lambda_2}} \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}^*_{2-k})(\boldsymbol{Y}'_0\boldsymbol{Y}_0 + \lambda_2\boldsymbol{I})\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}^*_{2-k}) + \frac{\lambda_1}{2}\boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}^*_{2-k})\boldsymbol{Y}'_0\boldsymbol{y}_T$$

for $k \in \{0,1\}$. Similarly, for VT regression, we proceed as above to obtain

$$\widehat{Y}^{\mathrm{vt}}_{NT}(0) = \frac{\langle \boldsymbol{y}_T, \widehat{\boldsymbol{\beta}}^*_1 \circ \widehat{\boldsymbol{\beta}}^*_2 \rangle}{\sqrt{1+\lambda_2}}, \tag{34}$$

where

$$\widehat{\boldsymbol{\beta}}^*_{1+k} = \left(\frac{1}{\sqrt{1+\lambda_2}} \boldsymbol{D}(\widehat{\boldsymbol{\beta}}^*_{2-k})(\boldsymbol{Y}_0\boldsymbol{Y}'_0 + \lambda_2\boldsymbol{I})\boldsymbol{D}(\widehat{\boldsymbol{\beta}}^*_{2-k}) + \frac{\lambda_1}{2}\boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\beta}}^*_{2-k})\boldsymbol{Y}_0\boldsymbol{y}_N$$

for $k \in \{0,1\}$. Leveraging (33) and (34), we find that the HZ regression solution can be linear in y and at least quadratic in q. On the other hand, the VT regression solution can be linear in q and at least quadratic in y. Since the elastic net regression solution is unique, provided $\lambda_2 > 0$, this implies that HZ and VT regressions do not yield matching solutions in general. *Q.E.D.*

C.6.6. *Proof of Lemma 10: Simplex Regression*

PROOF: Consider HZ regression. We write the Lagrangian of (4) as

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}} \ \|\boldsymbol{y}_T - \boldsymbol{Y}_0\boldsymbol{\alpha}\|^2_2 \ + \ \lambda\|\boldsymbol{\alpha}\|^2_2 \ - \ (\boldsymbol{\theta}^{\mathrm{hz}})'\boldsymbol{\alpha} + \nu^{\mathrm{hz}}(\boldsymbol{1}'\boldsymbol{\alpha} - 1),$$

where $\boldsymbol{\theta}^{\mathrm{hz}} \in \mathbb{R}^{T_0}$ and $\nu^{\mathrm{hz}} \in \mathbb{R}$. By the Karush-Kuhn-Tucker (KKT) conditions, optimality is achieved if the following are satisfied:

$$\widehat{\boldsymbol{\alpha}} \succeq \boldsymbol{0}, \quad \boldsymbol{1}'\widehat{\boldsymbol{\alpha}} = 1,$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{hz}} \succeq \boldsymbol{0},$$

$$\widehat{\theta}^{\mathrm{hz}}_i\widehat{\alpha}_i = 0 \quad \text{for } i = 1, \ldots, T_0,$$

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Y}'_0\boldsymbol{Y}_0 + \lambda\boldsymbol{I})^{-1} \left(\boldsymbol{Y}'_0\boldsymbol{y}_T + \frac{1}{2}\widehat{\boldsymbol{\theta}}^{\mathrm{hz}} - \frac{\widehat{\nu}^{\mathrm{hz}}}{2}\boldsymbol{1} \right).$$

Therefore, given primal and dual feasible variables $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\theta}}^{\mathrm{hz}}, \widehat{\nu}^{\mathrm{hz}})$, we can write the final HZ prediction as

$$\widehat{Y}^{\mathrm{hz}}_{NT}(0) = \widehat{Y}^{\mathrm{hz,ols}}_{NT}(0) + (1/2)\boldsymbol{y}'_N(\boldsymbol{Y}'_0\boldsymbol{Y}_0 + \lambda\boldsymbol{I})^{-1}(\widehat{\boldsymbol{\theta}}^{\mathrm{hz}} - \widehat{\nu}^{\mathrm{hz}}\boldsymbol{1}),$$

where $\widehat{Y}^{\mathrm{hz,ols}}_{NT}(0) = \boldsymbol{y}'_N(\boldsymbol{Y}'_0\boldsymbol{Y}_0 + \lambda\boldsymbol{I})^{-1}\boldsymbol{Y}'_0\boldsymbol{y}_T$ converges to the prediction corresponding to the OLS solution with minimum ℓ_2-norm as $\lambda \to 0^+$. Similarly, for VT regression, the KKT conditions are

$$\widehat{\boldsymbol{\beta}} \succeq \boldsymbol{0}, \quad \boldsymbol{1}'\widehat{\boldsymbol{\beta}} = 1,$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} \succeq \mathbf{0},$$

$$\widehat{\theta}_i^{\mathrm{vt}} \widehat{\beta}_i = 0 \quad \text{for } i = 1, \ldots, N_0,$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda \boldsymbol{I})^{-1} \left(\boldsymbol{Y}_0 \boldsymbol{y}_N + \frac{1}{2} \widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \frac{\widehat{\nu}^{\mathrm{vt}}}{2} \mathbf{1} \right).$$

For primal and dual feasible variables $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}^{\mathrm{vt}}, \widehat{\nu}^{\mathrm{vt}})$, this yields

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \widehat{Y}_{NT}^{\mathrm{vt,ols}}(0) + (1/2) \boldsymbol{y}_T' (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda \boldsymbol{I})^{-1} (\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1}),$$

where $\widehat{Y}_{NT}^{\mathrm{vt,ols}}(0) = \boldsymbol{y}_T' (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda \boldsymbol{I})^{-1} \boldsymbol{Y}_0 \boldsymbol{y}_N$ converges to the prediction corresponding to the OLS solution with minimum ℓ_2-norm as $\lambda \to 0^+$. Notably, as per Theorem 1, $\widehat{Y}_{NT}^{\mathrm{hz,ols}}(0) = \widehat{Y}_{NT}^{\mathrm{vt,ols}}(0) = \widehat{Y}_{NT}^{\mathrm{ols}}(0)$ for any $\lambda \geq 0$. As a result,

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = \widehat{Y}_{NT}^{\mathrm{ols}}(0) + (1/2) \boldsymbol{y}_N' (\boldsymbol{Y}_0' \boldsymbol{Y}_0 + \lambda \boldsymbol{I})^{-1} (\widehat{\boldsymbol{\theta}}^{\mathrm{hz}} - \widehat{\nu}^{\mathrm{hz}} \mathbf{1}) \tag{35}$$

$$\widehat{Y}_{NT}^{\mathrm{vt}}(0) = \widehat{Y}_{NT}^{\mathrm{ols}}(0) + (1/2) \boldsymbol{y}_T' (\boldsymbol{Y}_0 \boldsymbol{Y}_0' + \lambda \boldsymbol{I})^{-1} (\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1}). \tag{36}$$

As seen from (35) and (36), the leading terms in the HZ and VT simplex regression predictions are identical. The remaining terms, however, can differ from one another. As an example, consider $N = T$ with

$$\boldsymbol{Y}_0 = \boldsymbol{I}, \quad \boldsymbol{y}_N = \mathbf{0}, \quad \boldsymbol{y}_T = (1 + \lambda)(\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1}). \tag{37}$$

By construction, observe that

$$\widehat{\boldsymbol{\beta}} = \frac{1}{2(1 + \lambda)} (\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1}).$$

Recall from the KKT conditions for VT regression that $\widehat{\boldsymbol{\beta}} \succeq \mathbf{0}$ and $\mathbf{1}' \widehat{\boldsymbol{\beta}} = 1$. Therefore, at least one entry of $(\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1})$ must be strictly positive. This yields

$$(1 + \lambda)^{-1} \boldsymbol{y}_T' (\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1}) = (\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1})' (\widehat{\boldsymbol{\theta}}^{\mathrm{vt}} - \widehat{\nu}^{\mathrm{vt}} \mathbf{1}) > 0. \tag{38}$$

Plugging (37) and (38) into (35) and (36), we obtain

$$\widehat{Y}_{NT}^{\mathrm{hz}}(0) = 0 \quad \text{and} \quad \widehat{Y}_{NT}^{\mathrm{vt}}(0) > 0,$$

which concludes our proof. *Q.E.D.*

C.6.7. *Proof of Lemma 11: ℓ_p-penalties*

PROOF: We recall the Hadamard product parametrization (HPP): for any vector $\boldsymbol{z}$ and integer $K \geq 1$,

$$\|\boldsymbol{z}\|_p^p = \min_{\boldsymbol{z}_1 \circ \cdots \circ \boldsymbol{z}_K = \boldsymbol{z}} \frac{1}{K} \sum_{k=1}^K \|\boldsymbol{z}_k\|_2^2,$$

where $\circ$ denotes the Hadamard (componentwise) product. We rewrite our subclass of ℓ_p-penalties, i.e., (28) and (29), as sums of ℓ_2-penalties via the HPP technique:

$$(\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_K) = \operatorname*{argmin}_{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K} \|\boldsymbol{y}_T - \boldsymbol{Y}_0(\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K)\|_2^2 \; + \; \frac{\lambda}{K} \sum_{k=1}^{K} \|\boldsymbol{\alpha}_k\|_2^2 \qquad (39)$$

$$(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K) = \operatorname*{argmin}_{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K} \|\boldsymbol{y}_N - \boldsymbol{Y}_0'(\boldsymbol{\beta}_1 \circ \cdots \circ \boldsymbol{\beta}_K)\|_2^2 \; + \; \frac{\lambda}{K} \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_2^2, \qquad (40)$$

where $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}_1 \circ \cdots \circ \widehat{\boldsymbol{\alpha}}_K$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_1 \circ \cdots \circ \widehat{\boldsymbol{\beta}}_K$. Below, we leverage the results of Hoff (2017), which provides an alternating ridge regression algorithm to solve for (39)–(40).

Consider HZ regression. Let us solve for $\boldsymbol{\alpha}_k$ for $k \in [K]$ by fixing $\boldsymbol{\alpha}_{k'}$ for $k' \neq k$. By the optimality conditions,

$$\nabla_{\boldsymbol{\alpha}_k} \left\{ (\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K)' \boldsymbol{Y}_0' \boldsymbol{Y}_0 (\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K) - 2(\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K)' \boldsymbol{Y}_0' \boldsymbol{y}_T + \frac{\lambda}{K} \boldsymbol{\alpha}_k' \boldsymbol{\alpha}_k \right\} = 0. \qquad (41)$$

In order to solve for (41), observe that

$$(\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K)' \boldsymbol{Y}_0' \boldsymbol{Y}_0 (\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K) = \boldsymbol{\alpha}_k' (\boldsymbol{Y}_0' \boldsymbol{Y}_0 \circ \boldsymbol{\alpha}_{\sim k} \boldsymbol{\alpha}_{\sim k}') \boldsymbol{\alpha}_k$$

$$(\boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_K)' \boldsymbol{Y}_0' \boldsymbol{y}_T = \boldsymbol{\alpha}_k' (\boldsymbol{\alpha}_{\sim k} \circ \boldsymbol{Y}_0' \boldsymbol{y}_T),$$

where $\boldsymbol{\alpha}_{\sim k} = \boldsymbol{\alpha}_1 \circ \cdots \circ \boldsymbol{\alpha}_{k-1} \circ \boldsymbol{\alpha}_{k+1} \circ \cdots \circ \boldsymbol{\alpha}_K$. This allows us to rewrite (41) as

$$\nabla_{\boldsymbol{\alpha}_k} \left\{ \boldsymbol{\alpha}_k' \left(\boldsymbol{Y}_0' \boldsymbol{Y}_0 \circ \boldsymbol{\alpha}_{\sim k} \boldsymbol{\alpha}_{\sim k}' + \frac{\lambda}{K} \boldsymbol{I} \right) \boldsymbol{\alpha}_k - 2 \boldsymbol{\alpha}_k' (\boldsymbol{\alpha}_{\sim k} \circ \boldsymbol{Y}_0' \boldsymbol{y}_T) \right\} = 0.$$

This is quadratic in $\boldsymbol{\alpha}_k$ for fixed $\boldsymbol{\alpha}_{\sim k}$. Thus, the unique minimizer at convergence is

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{Y}_0' \boldsymbol{Y}_0 \circ \widehat{\boldsymbol{\alpha}}_{\sim k} \widehat{\boldsymbol{\alpha}}_{\sim k}' + \frac{\lambda}{K} \boldsymbol{I} \right)^{-1} (\widehat{\boldsymbol{\alpha}}_{\sim k} \circ \boldsymbol{Y}_0' \boldsymbol{y}_T), \qquad (42)$$

where $\widehat{\boldsymbol{\alpha}}_{\sim k} = \widehat{\boldsymbol{\alpha}}_1 \circ \cdots \circ \widehat{\boldsymbol{\alpha}}_{k-1} \circ \widehat{\boldsymbol{\alpha}}_{k+1} \circ \cdots \circ \widehat{\boldsymbol{\alpha}}_K$. Leveraging properties of the Hadamard product noted in Styan (1973), we rewrite

$$\boldsymbol{Y}_0' \boldsymbol{Y}_0 \circ \widehat{\boldsymbol{\alpha}}_{\sim k} \widehat{\boldsymbol{\alpha}}_{\sim k}' = \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k}) \boldsymbol{Y}_0' \boldsymbol{Y}_0 \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k})$$

$$\boldsymbol{Y}_0' \boldsymbol{y}_T \circ \widehat{\boldsymbol{\alpha}}_{\sim k} = \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k}) \boldsymbol{Y}_0' \boldsymbol{y}_T,$$

where $\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k})$ is the diagonal matrix formed from $\widehat{\boldsymbol{\alpha}}_{\sim k}$. Leveraging these equalities, we simplify (42) as

$$\widehat{\boldsymbol{\alpha}}_k = \left(\boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k}) \boldsymbol{Y}_0' \boldsymbol{Y}_0 \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k}) + \frac{\lambda}{K} \boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\alpha}}_{\sim k}) \boldsymbol{Y}_0' \boldsymbol{y}_T.$$

We now turn to VT regression. Following the arguments above, for every $k \in [K]$,

$$\widehat{\boldsymbol{\beta}}_k = \left(\boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k}) \boldsymbol{Y}_0 \boldsymbol{Y}_0' \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k}) + \frac{\lambda}{K} \boldsymbol{I} \right)^{-1} \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k}) \boldsymbol{Y}_0 \boldsymbol{y}_N,$$

where $\widehat{\boldsymbol{\beta}}_{\sim k} = \widehat{\boldsymbol{\beta}}_1 \circ \cdots \circ \widehat{\boldsymbol{\beta}}_{k-1} \circ \widehat{\boldsymbol{\beta}}_{k+1} \circ \cdots \circ \widehat{\boldsymbol{\beta}}_K$ and $\boldsymbol{D}(\widehat{\boldsymbol{\beta}}_{\sim k})$ is the diagonal matrix formed from $\widehat{\boldsymbol{\beta}}_{\sim k}$. This completes the proof. *Q.E.D.*

APPENDIX D: Proofs for Inference

D.1. Proof of Theorem 3

To establish Theorem 3, we first state a few useful results.

LEMMA 12—Theorem 2.7.1 of Lehmann (2000): *Let X_i for $i = 1, \ldots, n$ be independently distributed with means $\mathbb{E}[X_i] = \zeta_i$ and variances σ_i^2, and with finite third moments. Let $\bar{X} = (1/n)\sum_{i=1}^n X_i$. Then*

$$\frac{\bar{X} - \mathbb{E}[\bar{X}]}{\mathrm{Var}(\bar{X})^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1),$$

provided

$$\left(\sum_{i=1}^n \mathbb{E}\left[|X_i - \zeta_i|^3 \right] \right)^2 = o\left(\left(\sum_{i=1}^n \sigma_i^2 \right)^3 \right).$$

LEMMA 13: *Consider a random vector $\boldsymbol{x}$ and random matrix $\boldsymbol{A}$. Let $\mathbb{E}[\boldsymbol{x}|\boldsymbol{A}] = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{x}|\boldsymbol{A}) = \boldsymbol{\Sigma}$. Then $\mathbb{E}[\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}|\boldsymbol{A}] = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma})$.*

PROOF: (i) [HZ model] Let Assumptions 1–2 hold. By (21), Lemma 12 yields

$$\frac{\widehat{Y}_{NT}(0) - \mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0]}{\mathrm{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0)^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

To evaluate $\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0]$, we first observe that

$$\begin{aligned}
\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0] &= \mathbb{E}[\langle \boldsymbol{y}_N, \widehat{\boldsymbol{\alpha}} \rangle | \boldsymbol{y}_N, \boldsymbol{Y}_0] \\
&= \mathbb{E}[\langle \boldsymbol{y}_N, \boldsymbol{Y}_0^\dagger \boldsymbol{y}_T \rangle | \boldsymbol{y}_N, \boldsymbol{Y}_0] \\
&= \boldsymbol{y}_N' \boldsymbol{Y}_0^\dagger \mathbb{E}[\boldsymbol{y}_T | \boldsymbol{y}_N, \boldsymbol{Y}_0] \\
&= \boldsymbol{y}_N' \boldsymbol{Y}_0^\dagger \boldsymbol{Y}_0 \boldsymbol{\alpha}^* \\
&= \boldsymbol{y}_N' \boldsymbol{H}^v \boldsymbol{\alpha}^*.
\end{aligned} \tag{43}$$

Moving to the variance term, we note that

$$\mathrm{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0) = \boldsymbol{y}_N' \, \mathrm{Cov}(\widehat{\boldsymbol{\alpha}}|\boldsymbol{y}_N, \boldsymbol{Y}_0)\boldsymbol{y}_N. \tag{44}$$

Towards evaluating the above, we note that

$$\begin{aligned}
\mathrm{Cov}(\widehat{\boldsymbol{\alpha}}|\boldsymbol{y}_N, \boldsymbol{Y}_0) &= \mathrm{Cov}(\boldsymbol{Y}_0^\dagger \boldsymbol{y}_T | \boldsymbol{y}_N, \boldsymbol{Y}_0) \\
&= \boldsymbol{Y}_0^\dagger \, \mathrm{Cov}(\boldsymbol{y}_T | \boldsymbol{y}_N, \boldsymbol{Y}_0)(\boldsymbol{Y}_0')^\dagger \\
&= \boldsymbol{Y}_0^\dagger \, \mathrm{Cov}(\boldsymbol{\varepsilon}_T | \boldsymbol{y}_N, \boldsymbol{Y}_0)(\boldsymbol{Y}_0')^\dagger \\
&= \boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\mathrm{hz}}(\boldsymbol{Y}_0')^\dagger.
\end{aligned} \tag{45}$$

Plugging (45) into (44), we obtain

$$\text{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{y}_N, \boldsymbol{Y}_0) = \boldsymbol{y}'_N \boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\text{hz}} (\boldsymbol{Y}'_0)^\dagger \boldsymbol{y}_N = \widehat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_T^{\text{hz}} \widehat{\boldsymbol{\beta}}, \tag{46}$$

where we recall that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Y}'_0)^\dagger \boldsymbol{y}_N$. Putting it all together, we conclude

$$\frac{\widehat{Y}_{NT}(0) - \langle \boldsymbol{y}_N, \boldsymbol{H}^v \boldsymbol{\alpha}^* \rangle}{(\widehat{\boldsymbol{\beta}}' \boldsymbol{\Sigma}_T^{\text{hz}} \widehat{\boldsymbol{\beta}})^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

(ii) [VT model] Let Assumptions 3–4 hold. Following the arguments above, we have

$$\frac{\widehat{Y}_{NT}(0) - \langle \boldsymbol{y}_T, \boldsymbol{H}^u \boldsymbol{\beta}^* \rangle}{(\widehat{\boldsymbol{\alpha}}' \boldsymbol{\Sigma}_N^{\text{vt}} \widehat{\boldsymbol{\alpha}})^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

(iii) [Mixed model] Let Assumptions 5–6 hold. We will find it useful to write

$$\widehat{Y}_{NT}(0) = \boldsymbol{y}'_N \boldsymbol{Y}_0^\dagger \boldsymbol{y}_T = \sum_{i \le N_0} \sum_{t \le T_0} (\boldsymbol{Y}_0^\dagger)_{it} Y_{iT} Y_{Nt}. \tag{47}$$

By Assumption 6, (47) is a sum of independent random variables with

$$\mathbb{E}[Y_{iT} Y_{Nt}|\boldsymbol{Y}_0] = \mathbb{E}[Y_{iT}|\boldsymbol{Y}_0]\mathbb{E}[Y_{Nt}|\boldsymbol{Y}_0]$$

$$\text{Var}(Y_{iT} Y_{Nt}|\boldsymbol{Y}_0) = \mathbb{E}[Y_{iT}|\boldsymbol{Y}_0]^2 \sigma_{Nt}^2 + \mathbb{E}[Y_{Nt}|\boldsymbol{Y}_0]^2 \sigma_{iT}^2 + \sigma_{iT}^2 \sigma_{Nt}^2.$$

Lemma 12 then establishes that

$$\frac{\widehat{Y}_{NT}(0) - \mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0]}{\text{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0)^{1/2}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Our aim is to evaluate $\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0]$ and $\text{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0)$. Towards the former, we use Assumptions 5–6 with the law of total expectation to obtain

$$\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0] = \mathbb{E}\left[\mathbb{E}[\langle \boldsymbol{y}_N, \boldsymbol{Y}_0^\dagger \boldsymbol{y}_T \rangle|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0]|\boldsymbol{Y}_0\right]$$

$$= \mathbb{E}\left[\mathbb{E}[\boldsymbol{y}'_N \boldsymbol{Y}_0^\dagger (\boldsymbol{Y}_0 \boldsymbol{\alpha}^* + \boldsymbol{\varepsilon}_T)|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0]|\boldsymbol{Y}_0\right]$$

$$= \mathbb{E}\left[(\boldsymbol{Y}'_0 \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_N)' \boldsymbol{Y}_0^\dagger \boldsymbol{Y}_0 \boldsymbol{\alpha}^*|\boldsymbol{Y}_0\right]$$

$$= \langle \boldsymbol{\beta}^*, \boldsymbol{Y}_0 \boldsymbol{\alpha}^* \rangle.$$

Note that we have used the fact that $\boldsymbol{y}_N$ is deterministic given $(\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0)$. Similarly, by the law of total variance,

$$\text{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{Y}_0) = \mathbb{E}[\text{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0)|\boldsymbol{Y}_0] + \text{Var}(\mathbb{E}[\widehat{Y}_{NT}(0)|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0]|\boldsymbol{Y}_0). \tag{48}$$

Following the derivation of (46), we have

$$\mathbb{E}[\text{Var}(\widehat{Y}_{NT}(0)|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0)|\boldsymbol{Y}_0]$$

$$= \mathbb{E}[\boldsymbol{y}_N' \boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{y}_N | \boldsymbol{Y}_0]$$

$$= (\boldsymbol{Y}_0' \boldsymbol{\beta}^*)' \boldsymbol{A} (\boldsymbol{Y}_0' \boldsymbol{\beta}^*) + \mathbb{E}[\boldsymbol{\varepsilon}_N' \boldsymbol{A} \boldsymbol{\varepsilon}_N | \boldsymbol{Y}_0] + 2\mathbb{E}[\boldsymbol{\varepsilon}_N' \boldsymbol{Y}_0' \boldsymbol{\beta}^* | \boldsymbol{Y}_0], \qquad (49)$$

where $\boldsymbol{A} = \boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{Y}_0')^\dagger$. Notice that Assumption 6 gives $\mathbb{E}[\boldsymbol{\varepsilon}_N' \boldsymbol{Y}_0' \boldsymbol{\beta}^* | \boldsymbol{Y}_0] = 0$. Since $\boldsymbol{A}$ is deterministic given $\boldsymbol{Y}_0$, Lemma 13 yields

$$\mathbb{E}[\boldsymbol{\varepsilon}_N' \boldsymbol{A} \boldsymbol{\varepsilon}_N | \boldsymbol{Y}_0] = \text{tr}(\boldsymbol{A} \boldsymbol{\Sigma}_N^{\text{mix}}). \qquad (50)$$

Following the arguments that led to the derivation of (43), we have

$$\text{Var}(\mathbb{E}[\widehat{Y}_{NT}(0) | \boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0] | \boldsymbol{Y}_0) = \text{Var}(\boldsymbol{y}_N' \boldsymbol{H}^v \boldsymbol{\alpha}^* | \boldsymbol{Y}_0) = (\boldsymbol{H}^v \boldsymbol{\alpha}^*)' \boldsymbol{\Sigma}_N^{\text{mix}} (\boldsymbol{H}^v \boldsymbol{\alpha}^*). \qquad (51)$$

Plugging (49), (50), and (51) into (48), we arrive at

$$\text{Var}(\widehat{Y}_{NT}(0) | \boldsymbol{Y}_0)$$

$$= (\boldsymbol{H}^v \boldsymbol{\alpha}^*)' \boldsymbol{\Sigma}_N^{\text{mix}} (\boldsymbol{H}^v \boldsymbol{\alpha}^*) + (\boldsymbol{H}^u \boldsymbol{\beta}^*)' \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{H}^u \boldsymbol{\beta}^*) + \text{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\text{mix}}).$$

This completes the proof. *Q.E.D.*

D.2. Proofs for Model-Based Confidence Intervals

We first state a useful lemma to prove Lemmas 1–3.

LEMMA 14: [Mixed model] *Let Assumptions 5–6 hold. Then,*

$$\mathbb{E}[\widehat{v}_0^{\text{mix}} | \boldsymbol{Y}_0] = (\boldsymbol{H}^u \boldsymbol{\beta}^*)' \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T | \boldsymbol{Y}_0] (\boldsymbol{H}^u \boldsymbol{\beta}^*) + \text{tr}(\boldsymbol{Y}_0^\dagger \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T | \boldsymbol{Y}_0] (\boldsymbol{Y}_0')^\dagger \boldsymbol{\Sigma}_N^{\text{mix}})$$

$$+ (\boldsymbol{H}^v \boldsymbol{\alpha}^*)' \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N | \boldsymbol{Y}_0] (\boldsymbol{H}^v \boldsymbol{\alpha}^*) + \text{tr}(\boldsymbol{Y}_0^\dagger \boldsymbol{\Sigma}_T^{\text{mix}} (\boldsymbol{Y}_0')^\dagger \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N | \boldsymbol{Y}_0])$$

$$- \text{tr}(\boldsymbol{Y}_0^\dagger \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T | \boldsymbol{Y}_0] (\boldsymbol{Y}_0')^\dagger \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N | \boldsymbol{Y}_0]).$$

D.2.1. Proof of Lemma 1

PROOF: (i) [HZ model] Let Assumptions 1–2 hold. Taking note that $\boldsymbol{H}_\perp^u \boldsymbol{Y}_0 = \boldsymbol{0}$,

$$\|\boldsymbol{H}_\perp^u \boldsymbol{y}_T\|_2^2 = \boldsymbol{y}_T' \boldsymbol{H}_\perp^u \boldsymbol{y}_T$$

$$= (\boldsymbol{Y}_0 \boldsymbol{\alpha} + \boldsymbol{\varepsilon}_T)' \boldsymbol{H}_\perp^u (\boldsymbol{Y}_0 \boldsymbol{\alpha} + \boldsymbol{\varepsilon}_T)$$

$$= \boldsymbol{\varepsilon}_T' \boldsymbol{H}_\perp^u \boldsymbol{\varepsilon}_T.$$

Applying Lemma 13 then gives

$$\mathbb{E}[\boldsymbol{\varepsilon}_T' \boldsymbol{H}_\perp^u \boldsymbol{\varepsilon}_T | \boldsymbol{y}_N, \boldsymbol{Y}_0] = \text{tr}(\boldsymbol{H}_\perp^u)(\sigma_T^{\text{hz}})^2 = (N_0 - R)(\sigma_T^{\text{hz}})^2, \qquad (52)$$

where the final equality follows because the trace of a projection matrix equals its rank. Taken altogether, we have $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{homo}} | \boldsymbol{y}_N, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\text{hz}}$. Therefore,

$$\mathbb{E}[\widehat{v}_0^{\text{hz,homo}} | \boldsymbol{y}_N, \boldsymbol{Y}_0] = \widehat{\boldsymbol{\beta}}' \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{homo}} | \boldsymbol{y}_N, \boldsymbol{Y}_0] \widehat{\boldsymbol{\beta}} = v_0^{\text{hz}}.$$

(ii) [VT model] Let Assumptions 3–4 hold. Following the arguments above, we conclude that $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\text{homo}} | \boldsymbol{y}_T, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\text{vt}}$ and $\mathbb{E}[\widehat{v}_0^{\text{vt,homo}} | \boldsymbol{y}_T, \boldsymbol{Y}_0] = v_0^{\text{vt}}$.

(iii) [Mixed model] Let Assumptions 5–6 hold. Following the arguments that led to (52), we obtain $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{homo}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\mathrm{mix}}$ and $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{homo}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{mix}}$. Applying Lemma 14 then gives $\mathbb{E}[\widehat{v}_0^{\mathrm{mix,homo}}|\boldsymbol{Y}_0] = v_0^{\mathrm{mix}}$. The proof is complete. *Q.E.D.*

D.3. *Proof of Lemma 2*

PROOF: Before we establish the biases of $(\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{jack}}, \widehat{\boldsymbol{\Sigma}}_N^{\mathrm{jack}})$, we first justify their forms. As noted in Section 4.1.3, jackknife is a popular approach to estimate the covariances of $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$. Below, we follow the standard techniques to derive the jackknife estimate of these objects, which will then be used to derive $(\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{jack}}, \widehat{\boldsymbol{\Sigma}}_N^{\mathrm{jack}})$. Without loss of generality, we begin with $\widehat{\boldsymbol{\alpha}}$. Notably, while standard derivations consider $\boldsymbol{Y}_0$ with full column rank, we consider a general matrix $\boldsymbol{Y}_0$ that may be rank deficient. This difference is subtle so the following proof is by no means novel. We provide it simply for completeness.

To describe the jackknife, we define $\widehat{\boldsymbol{\alpha}}_{\sim i}$ as the minimum ℓ_2-norm solution to (1), where $\lambda_1 = \lambda_2 = 0$, without the ith observation, i.e.,

$$\widehat{\boldsymbol{\alpha}}_{\sim i} = (\boldsymbol{Y}_{0,\sim i}' \boldsymbol{Y}_{0,\sim i})^\dagger \boldsymbol{Y}_{0,\sim i}' \boldsymbol{y}_{T,\sim i}, \tag{53}$$

where $\boldsymbol{Y}_{0,\sim i}$ and $\boldsymbol{y}_{T,\sim i}$ correspond to $\boldsymbol{Y}_0$ and $\boldsymbol{y}_T$ without the ith observation. We define the pseudo-estimator as $\tilde{\boldsymbol{\alpha}}_i = T_0 \widehat{\boldsymbol{\alpha}} - (T_0 - 1)\widehat{\boldsymbol{\alpha}}_{\sim i}$. With these quantities defined, we write the jackknife variance estimator as

$$\widehat{\boldsymbol{V}}^{\mathrm{jack}} = \frac{1}{(T_0 - 1)^2} \sum_{i \leq N_0} (\tilde{\boldsymbol{\alpha}}_i - \widehat{\boldsymbol{\alpha}})(\tilde{\boldsymbol{\alpha}}_i - \widehat{\boldsymbol{\alpha}})'. \tag{54}$$

To evaluate this quantity, we will rewrite $\widehat{\boldsymbol{\alpha}}_{\sim i}$ in a more convenient form. In particular,

$$\boldsymbol{Y}_{0,\sim i}' \boldsymbol{Y}_{0,\sim i} = \boldsymbol{Y}_0' \boldsymbol{Y}_0' - \boldsymbol{y}_i \boldsymbol{y}_i'$$
$$\boldsymbol{Y}_{0,\sim i}' \boldsymbol{y}_{T,\sim i} = \boldsymbol{Y}_0' \boldsymbol{y}_T - \boldsymbol{y}_i Y_{iT},$$

where $\boldsymbol{y}_i = [Y_{it} : t \leq T_0]$ is the ith row of $\boldsymbol{Y}_0$. We do not assume that $\boldsymbol{Y}_0' \boldsymbol{Y}_0$ is nonsingular. As such, we use a generalized form of the Sherman-Morrison formula (Cline, 1965, Meyer, 1973) to obtain

$$(\boldsymbol{Y}_{0,\sim i}' \boldsymbol{Y}_{0,\sim i})^\dagger = (\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger + (1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i \boldsymbol{y}_i' (\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger. \tag{55}$$

Recall $\widehat{\boldsymbol{\alpha}} = (\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{Y}_0' \boldsymbol{y}_T$ and note $Y_{iT} - \boldsymbol{y}_i' \widehat{\boldsymbol{\alpha}}$ is the ith element of $\widehat{\boldsymbol{\varepsilon}}_T = \boldsymbol{H}_\perp^u \boldsymbol{y}_T$. Using these facts, we plug (55) into (53) to yield

$$\widehat{\boldsymbol{\alpha}}_{\sim i} = \left[(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger + (1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i \boldsymbol{y}_i' (\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \right] (\boldsymbol{Y}_0' \boldsymbol{y}_T - \boldsymbol{y}_i Y_{iT})$$
$$= \widehat{\boldsymbol{\alpha}} - (\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i Y_{iT} + (1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i \boldsymbol{y}_i' \widehat{\boldsymbol{\alpha}} - H_{ii}^u(1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i Y_{iT}$$
$$= \widehat{\boldsymbol{\alpha}} - (1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i \widehat{\varepsilon}_{iT}. \tag{56}$$

Inserting (56) into our pseudo-estimate, we have

$$\tilde{\boldsymbol{\alpha}}_i = T_0 \widehat{\boldsymbol{\alpha}} - (T_0 - 1)\left(\widehat{\boldsymbol{\alpha}} - (1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i \widehat{\varepsilon}_{iT} \right)$$
$$= \widehat{\boldsymbol{\alpha}} + (T_0 - 1)(1 - H_{ii}^u)^{-1}(\boldsymbol{Y}_0' \boldsymbol{Y}_0)^\dagger \boldsymbol{y}_i \widehat{\varepsilon}_{iT}. \tag{57}$$

Inserting (57) into (54), we have

$$\widehat{\boldsymbol{V}}^{\text{jack}} = (\boldsymbol{Y}_0'\boldsymbol{Y}_0)^\dagger \left(\sum_{i \leq N_0} \frac{\widehat{\varepsilon}_{iT}^2}{(1 - H_{ii}^u)^2} \boldsymbol{y}_i \boldsymbol{y}_i' \right) (\boldsymbol{Y}_0'\boldsymbol{Y}_0)^\dagger$$

$$= (\boldsymbol{Y}_0'\boldsymbol{Y}_0)^\dagger \boldsymbol{Y}_0' \boldsymbol{\Omega} \boldsymbol{Y}_0 (\boldsymbol{Y}_0'\boldsymbol{Y}_0)^\dagger,$$

where $\boldsymbol{\Omega}$ is a diagonal matrix with $\Omega_{ii} = \widehat{\varepsilon}_{iT}^2 (1 - H_{ii}^u)^{-2}$. Equivalently, $\boldsymbol{\Omega} = \text{diag}([\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I}]^\dagger [\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T])$. It then follows that

$$\boldsymbol{y}_N' \widehat{\boldsymbol{V}}^{\text{jack}} \boldsymbol{y}_N = \widehat{\boldsymbol{\beta}}' \boldsymbol{\Omega} \widehat{\boldsymbol{\beta}}.$$

To arrive at (18), we define $\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}} = \boldsymbol{\Omega}$. This corresponds to the EHW estimator with the jackknife correction. We derive (19) for $\widehat{\boldsymbol{\beta}}$ by applying the same arguments above. Now, we will evaluate the biases of $(\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}}, \widehat{\boldsymbol{\Sigma}}_N^{\text{jack}})$.

   (i) [HZ model] Let Assumptions 1–2 hold. We define $(\sigma_{iT}^{\text{hz}})^2 = \text{Var}(\varepsilon_{iT}|\boldsymbol{y}_N, \boldsymbol{Y}_0)$ for $i = 1, \ldots, N_0$. Observe that

$$\mathbb{E}[(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})^\dagger (\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T)|\boldsymbol{y}_N, \boldsymbol{Y}_0] = (\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})^\dagger \mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0]. \quad (58)$$

To evaluate (58), we follow the derivations of (43) and (45) to obtain

$$\mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \boldsymbol{H}_\perp^u \boldsymbol{Y}_0 \boldsymbol{\alpha}^* = \boldsymbol{0} \quad (59)$$

$$\text{Cov}(\widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0) = \boldsymbol{H}_\perp^u \boldsymbol{\Sigma}_T^{\text{hz}} \boldsymbol{H}_\perp^u. \quad (60)$$

Recall that $\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2$ for any random variable X. Thus, combining (59) with (60) gives

$$\mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0] = (\boldsymbol{H}_\perp^u \boldsymbol{\Sigma}_T^{\text{hz}} \boldsymbol{H}_\perp^u \circ \boldsymbol{I})\boldsymbol{1}. \quad (61)$$

Let $\widehat{\boldsymbol{\gamma}} = \mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0]$. By (61), the ℓth entry of $\widehat{\boldsymbol{\gamma}}$ can be written as

$$\widehat{\gamma}_\ell = \sum_{j \neq \ell} (H_{j\ell}^u)^2 (\sigma_{jT}^{\text{hz}})^2 + (1 - H_{\ell\ell}^u)^2 (\sigma_{\ell T}^{\text{hz}})^2,$$

where $H_{j\ell}^u$ is the (j, ℓ)th entry of $\boldsymbol{H}^u$. In turn, this allows us to rewrite (61) as

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u) \boldsymbol{\Sigma}_T^{\text{hz}} \boldsymbol{1}. \quad (62)$$

Next, let $\widehat{\boldsymbol{\zeta}} = (\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})^{-1} \widehat{\boldsymbol{\gamma}}$. Notice that the ℓth entry of $\widehat{\boldsymbol{\zeta}}$ is given by

$$\widehat{\zeta}_\ell = (\sigma_{\ell T}^{\text{hz}})^2 + \sum_{j \neq \ell} \frac{(H_{\ell j}^u)^2}{(1 - H_{\ell\ell}^u)^2} (\sigma_{jT}^{\text{hz}})^2.$$

Therefore, $\text{diag}(\widehat{\boldsymbol{\zeta}}) = \boldsymbol{\Sigma}_T^{\text{hz}} + \boldsymbol{\Delta}^{\text{hz}}$, where $\Delta_{\ell\ell}^{\text{hz}} = \sum_{j \neq \ell} (\sigma_{jT}^{\text{hz}})^2 (H_{\ell j}^u)^2 (1 - H_{\ell\ell}^u)^{-2}$ for $\ell = 1, \ldots, N_0$. Notice if $\max_\ell H_{\ell\ell}^u < 1$, then $(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})$ is nonsingular, i.e., the pseudo-inverse is precisely the inverse. In this situation, plugging the above into (58) gives

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\text{jack}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \text{diag}\left((\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})^{-1} \mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0]\right)$$

$$= \operatorname{diag}\left((\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u \circ \boldsymbol{I})^{-1}\widehat{\boldsymbol{\gamma}}\right)$$

$$= \operatorname{diag}(\widehat{\boldsymbol{\zeta}})$$

$$= \boldsymbol{\Sigma}_T^{\mathrm{hz}} + \boldsymbol{\Delta}^{\mathrm{hz}}. \tag{63}$$

From this, we conclude that

$$\mathbb{E}[\widehat{v}_0^{\mathrm{hz,jack}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \widehat{\boldsymbol{\beta}}'\, \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{jack}}|\boldsymbol{y}_N, \boldsymbol{Y}_0]\widehat{\boldsymbol{\beta}}$$

$$= \widehat{\boldsymbol{\beta}}'(\boldsymbol{\Sigma}_T^{\mathrm{hz}} + \boldsymbol{\Delta}^{\mathrm{hz}})\widehat{\boldsymbol{\beta}}$$

$$= v_0^{\mathrm{hz}} + \widehat{\boldsymbol{\beta}}'\boldsymbol{\Delta}^{\mathrm{hz}}\widehat{\boldsymbol{\beta}},$$

where we note that $\widehat{\boldsymbol{\beta}}'\boldsymbol{\Delta}^{\mathrm{hz}}\widehat{\boldsymbol{\beta}} \geq 0$.

(ii) [VT model] Let Assumptions 3–4 hold. Following the arguments above, we conclude $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{jack}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{vt}} + \boldsymbol{\Gamma}^{\mathrm{vt}}$, where $\Gamma_{\ell\ell}^{\mathrm{vt}} = \sum_{j\neq\ell}(\sigma_{Nj}^{\mathrm{vt}})^2(H_{\ell j}^v)^2(1 - H_{\ell\ell}^v)^{-2}$ for $\ell = 1, \ldots, T_0$. Thus, $\mathbb{E}[\widehat{v}_0^{\mathrm{vt,jack}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = v_0^{\mathrm{vt}} + \widehat{\boldsymbol{\alpha}}'\boldsymbol{\Gamma}^{\mathrm{vt}}\widehat{\boldsymbol{\alpha}}$, where we note that $\widehat{\boldsymbol{\alpha}}'\boldsymbol{\Gamma}^{\mathrm{vt}}\widehat{\boldsymbol{\alpha}} \geq 0$.

(ii) [Mixed model] Let Assumptions 5–6 hold. We define $(\sigma_{iT}^{\mathrm{mix}})^2 = \operatorname{Var}(\varepsilon_{iT}|\boldsymbol{Y}_0)$ for $i = 1, \ldots, N_0$ and $(\sigma_{Nt}^{\mathrm{mix}})^2 = \operatorname{Var}(\varepsilon_{Nt}|\boldsymbol{Y}_0)$ for $t = 1, \ldots, T_0$. Following the arguments that led to (63), we obtain $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{jack}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\mathrm{mix}} + \boldsymbol{\Delta}^{\mathrm{mix}}$, where $\Delta_{\ell\ell}^{\mathrm{mix}} = \sum_{j\neq\ell}(\sigma_{jT}^{\mathrm{mix}})^2(H_{\ell j}^u)^2(1 - H_{\ell\ell}^u)^{-2}$ for $\ell = 1, \ldots, N_0$. Similarly, we obtain $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{jack}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{mix}} + \boldsymbol{\Gamma}^{\mathrm{mix}}$, where $\Gamma_{\ell\ell}^{\mathrm{mix}} = \sum_{j\neq\ell}(\sigma_{Nj}^{\mathrm{mix}})^2(H_{\ell j}^v)^2(1 - H_{\ell\ell}^v)^{-2}$ for $\ell = 1, \ldots, T_0$. Applying Lemma 14 then gives

$$\mathbb{E}[\widehat{v}_0^{\mathrm{mix,jack}}|\boldsymbol{Y}_0]$$

$$= v_0^{\mathrm{mix}} + (\boldsymbol{H}^u\boldsymbol{\beta}^*)'\boldsymbol{\Delta}^{\mathrm{mix}}(\boldsymbol{H}^u\boldsymbol{\beta}^*) + (\boldsymbol{H}^v\boldsymbol{\alpha}^*)'\boldsymbol{\Gamma}^{\mathrm{mix}}(\boldsymbol{H}^v\boldsymbol{\alpha}^*) + \operatorname{tr}(\boldsymbol{Y}_0^\dagger\boldsymbol{\Delta}^{\mathrm{mix}}(\boldsymbol{Y}_0')^\dagger\boldsymbol{\Gamma}^{\mathrm{mix}}).$$

The proof is complete. _Q.E.D._

D.4. _Proof of Lemma 3_

PROOF: We adopt the strategy of Hartley et al. (1969) to prove our desired result.

(ii) [HZ model] Let Assumptions 1–2 hold. As in the proof of Lemma 2, we define $\widehat{\boldsymbol{\varepsilon}}_T = \boldsymbol{H}_\perp^u\boldsymbol{y}_T$. Observe

$$\mathbb{E}[(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)^{-1}(\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T)|\boldsymbol{y}_N, \boldsymbol{Y}_0] = (\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)^{-1}\mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0]. \tag{64}$$

To evaluate (64), we plug in (62) to obtain

$$\mathbb{E}[(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)^{-1}(\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T)|\boldsymbol{y}_N, \boldsymbol{Y}_0] = (\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)^{-1}(\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)\boldsymbol{\Sigma}_T^{\mathrm{hz}}\mathbf{1} = \boldsymbol{\Sigma}_T^{\mathrm{hz}}\mathbf{1}. \tag{65}$$

Plugging (65) into (64) yields

$$\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{HRK}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = \operatorname{diag}\left((\boldsymbol{H}_\perp^u \circ \boldsymbol{H}_\perp^u)^{-1}\mathbb{E}[\widehat{\boldsymbol{\varepsilon}}_T \circ \widehat{\boldsymbol{\varepsilon}}_T|\boldsymbol{y}_N, \boldsymbol{Y}_0]\right) = \boldsymbol{\Sigma}_T^{\mathrm{hz}}. \tag{66}$$

It then follows that $\mathbb{E}[\widehat{v}_0^{\mathrm{hz,HRK}}|\boldsymbol{y}_N, \boldsymbol{Y}_0] = v_0^{\mathrm{hz}}$.

(ii) [VT model] Let Assumptions 3–4 hold. Following the same arguments as above, we conclude $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{HRK}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{vt}}$ and $\mathbb{E}[\widehat{v}_0^{\mathrm{vt,HRK}}|\boldsymbol{y}_T, \boldsymbol{Y}_0] = v_0^{\mathrm{vt}}$.

(ii) [Mixed model] Let Assumptions 5–6 hold. Following the arguments that led to (66), we obtain $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T^{\mathrm{HRK}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_T^{\mathrm{mix}}$ and $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N^{\mathrm{HRK}}|\boldsymbol{Y}_0] = \boldsymbol{\Sigma}_N^{\mathrm{mix}}$. Applying Lemma 14 then gives $\mathbb{E}[\widehat{v}_0^{\mathrm{mix,HRK}}|\boldsymbol{Y}_0] = v_0^{\mathrm{mix}}$. The proof is complete.   *Q.E.D.*

D.5. *Proof of Lemma 14*

PROOF: By linearity of expectations,

$$\mathbb{E}[\widehat{v}_0^{\mathrm{mix}}|\boldsymbol{Y}_0] = \mathbb{E}[\widehat{v}_0^{\mathrm{hz}}|\boldsymbol{Y}_0] + \mathbb{E}[\widehat{v}_0^{\mathrm{vt}}|\boldsymbol{Y}_0] - \mathbb{E}[\mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\widehat{\boldsymbol{\Sigma}}_T(\boldsymbol{Y}_0')^{\dagger}\widehat{\boldsymbol{\Sigma}}_N)|\boldsymbol{Y}_0]. \tag{67}$$

We evaluate each term in (67).

Beginning with the first term, note that the randomness in $\widehat{\boldsymbol{\Sigma}}_T$ stems from $\boldsymbol{\varepsilon}_T$ and $\widehat{\boldsymbol{\beta}}$ is deterministic given $(\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0)$. As such, Assumptions 5–6 with Lemma 13 gives

$$\begin{aligned}
\mathbb{E}[\widehat{v}_0^{\mathrm{hz}}|\boldsymbol{Y}_0] &= \mathbb{E}[\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\Sigma}}_T\widehat{\boldsymbol{\beta}}|\boldsymbol{Y}_0] \\
&= \mathbb{E}\left[\mathbb{E}[\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\Sigma}}_T\widehat{\boldsymbol{\beta}}|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0]|\boldsymbol{Y}_0\right] \\
&= \mathbb{E}\left[\boldsymbol{y}_N'\boldsymbol{Y}_0^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T|\boldsymbol{Y}_0](\boldsymbol{Y}_0')^{\dagger}\boldsymbol{y}_N|\boldsymbol{Y}_0\right] \\
&= \mathbb{E}\left[(\boldsymbol{Y}_0'\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_N)\boldsymbol{Y}_0^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T|\boldsymbol{Y}_0](\boldsymbol{Y}_0')^{\dagger}(\boldsymbol{Y}_0'\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_N)|\boldsymbol{Y}_0\right] \\
&= (\boldsymbol{H}^u\boldsymbol{\beta}^*)'\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T|\boldsymbol{Y}_0](\boldsymbol{H}^u\boldsymbol{\beta}^*) + \mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T|\boldsymbol{Y}_0](\boldsymbol{Y}_0')^{\dagger}\boldsymbol{\Sigma}_N^{\mathrm{mix}}).
\end{aligned}$$

By an analogous argument, we derive

$$\mathbb{E}[\widehat{v}_0^{\mathrm{vt}}|\boldsymbol{Y}_0] = (\boldsymbol{H}^v\boldsymbol{\alpha}^*)'\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N|\boldsymbol{Y}_0](\boldsymbol{H}^v\boldsymbol{\alpha}^*) + \mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\boldsymbol{\Sigma}_T^{\mathrm{mix}}(\boldsymbol{Y}_0')^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N|\boldsymbol{Y}_0]).$$

Finally, we use the linearity of the trace operator with Assumption 6 to obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\widehat{\boldsymbol{\Sigma}}_T(\boldsymbol{Y}_0')^{\dagger}\widehat{\boldsymbol{\Sigma}}_N)|\boldsymbol{Y}_0] &= \mathbb{E}\left[\mathbb{E}[\mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\widehat{\boldsymbol{\Sigma}}_T(\boldsymbol{Y}_0')^{\dagger}\widehat{\boldsymbol{\Sigma}}_N)|\boldsymbol{\varepsilon}_N, \boldsymbol{Y}_0]|\boldsymbol{Y}_0\right] \\
&= \mathbb{E}\left[\mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T|\boldsymbol{Y}_0](\boldsymbol{Y}_0')^{\dagger}\widehat{\boldsymbol{\Sigma}}_N)|\boldsymbol{Y}_0\right] \\
&= \mathrm{tr}(\boldsymbol{Y}_0^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_T|\boldsymbol{Y}_0](\boldsymbol{Y}_0')^{\dagger}\mathbb{E}[\widehat{\boldsymbol{\Sigma}}_N|\boldsymbol{Y}_0]).
\end{aligned}$$

Putting everything together completes the proof.   *Q.E.D.*