

# Irrelevant Donors: Causes and Consequences of Spurious Pre-Treatment Fits in Causal Panel Methods

Frederico Guilherme Nogueira ([frederico.nogueira@gmail.com](mailto:frederico.nogueira@gmail.com))

## Abstract

Panel data methods for causal inference construct counterfactuals via linear combinations of control units, relying on relevance (treated unit's latent structure in donor span) and conditioning (stable recovery). Projection-based diagnostics fail to detect violations when irrelevant units contaminate pools. We formalize this via contamination ratio: irrelevant donors' effective rank over pre-treatment periods. Residuals attenuate proportionally; e.g., ratio 0.5 masks 50% of violations. Near unity under saturation, residuals vanish, yielding spurious fits with low pre-treatment error, a phase transition in projection-based diagnostic feasibility. Unstructured cases approximate rank to irrelevant donor count; structured irrelevance ties rank to confounding dimension, often smaller.

We prove that exact relevance is untestable using projection-based diagnostics due to the statistical contiguity of distributions under span membership and arbitrarily small violations, making them asymptotically indistinguishable. The minimum detectable violation diverges as contamination increases, with diagnostic power collapsing continuously toward saturation and the detection threshold inflating sharply. This provides the first explicit calibration of projection-based diagnostic feasibility as a function of contamination, sample size, and noise level.

Our results yield a comprehensive taxonomy classifying failures into refinable and irreducible categories: conditioning lapses, which lead to variance inflation and non-standard asymptotics, can be addressed via extended pre-treatment periods or regularization; whereas relevance failures induce structural, irreducible bias in dense factor environments, necessitating a fundamental rethink of research designs. These limits mandate prioritizing ex-ante donor curation guided by domain knowledge over ex-post algorithmic selection, as it is essential for maintaining diagnostic validity in practice. Furthermore, we motivate an alternative diagnostic framework in our companion work that avoids estimating the contamination ratio, weights, or factor dimensions, thereby bypassing the inherent circularity of residual-based approaches.

**Keywords:** Causal inference, panel data, identification, synthetic control, factor models, weak identification, diagnostic testing, impossibility results

## 1. Introduction

Panel data methods for causal inference, including synthetic controls (Abadie et al., 2010), difference-in-differences (Card & Krueger, 1994), and interactive fixed effects (Bai, 2009; Xu, 2017), aim to reconstruct untreated counterfactuals for treated units using weighted combinations of control donors. Their reliability hinges on whether the treated unit can be faithfully represented by the donors, a condition that, if violated, leads to biased estimates and invalid inference.

Standard practice assesses credibility via pre-treatment fit: low residuals are interpreted as evidence of valid identification. We show this diagnostic using projection-based diagnostics breaks down systematically when donor pools contain irrelevant units, those whose latent factors provide minimal signal about the treated unit. Methods can then produce seemingly robust results (excellent fit, stable weights) even when identification fails fundamentally.

## Identification primitives

We formalize identification through two geometric conditions:

**Relevance** requires the treated unit's latent loading  $\lambda_0$  lies in  $\text{span}(\Lambda)$ , ensuring  $\mathbb{E}[y_0] \in \text{span}\{\mathbb{E}[X]\}$ . When relevance fails, an orthogonal component induces structural bias no linear combination can recover.

**Conditioning** requires Gram matrices  $G_\Lambda = (1/N)\Lambda'\Lambda$  and  $G_F = (1/T_0)F'F$  have minimum eigenvalues bounded from zero, ensuring stable recovery. Weak conditioning produces variance inflation and nonstandard asymptotics, analogous to weak instruments (Staiger & Stock, 1997).

To clarify how partial geometric saturation can arise in empirically relevant settings, consider a scenario analogous to the Proposition 99 application in Abadie et al. (2010), which uses 38 U.S. states over a pre-treatment period of length  $T_0 = 19$ . Suppose the treated unit (California) loads on latent factors associated with technology-intensive industries, environmental regulation, and coastal services, while a subset of donor states is largely orthogonal to these dimensions. Examples include energy- and mining-dependent states driven by commodity price cycles, manufacturing-intensive Rust Belt states exposed to trade and industrial shocks, and Midwestern agricultural states primarily affected by weather and crop-yield fluctuations.

When such donors provide little signal about the treated unit's factor structure, they constitute a block of irrelevant controls in the sense of Definition 2.1. If the irrelevant donors are structured, sharing a small number of common orthogonal shocks within each group, then their contribution to the donor space is low-rank. For instance, with roughly three dominant factors per group, the effective rank of the irrelevant block is approximately  $s \approx 9$ , implying an effective saturation index  $\kappa = s/T_0 \approx 0.47$ . This illustrates that irrelevance alone is insufficient for projection-based diagnostic failure.

To distinguish this from the presence of irrelevance alone, note that low-rank structure among irrelevant donors keeps  $\kappa$  bounded away from 1 even with many such donors ( $N_{\text{irrel}} \gg T_0$ ), as their orthogonal subspace projection is confined to a low-dimensional manifold. In this case, diagnostics retain power, as the irrelevant block does not saturate the noise subspace, spurious alignments are limited to the shared orthogonal factors, leaving a detectable fraction of violations unmasked. Conversely, if irrelevant donors are fragmented (mutually uncorrelated or spanning multiple independent orthogonal bases), their effective rank approaches  $N_{\text{irrel}}$ , driving  $\kappa \rightarrow 1$  and fully saturating the space, even if each is economically irrelevant to the treated unit.

Our theory implies that in this structured irrelevance case, diagnostic power degrades proportionally to the effective saturation  $\kappa$ : nearly 47% of the structural violation induced by non-representability is absorbed by spurious geometric alignment, substantially weakening projection-based fit diagnostics even though the donor space remains far from fully saturated ( $\kappa < 1$ ). While this configuration is not meant to characterize the actual Proposition 99 application, it illustrates a general implication of the framework: moderate geometric saturation, arising from economically plausible low-rank heterogeneity among donors, can silently invalidate standard diagnostic checks even when the raw number of controls is modest.

The  $\kappa$ -mechanism. We formalize this diagnostic failure through a contamination model where irrelevant donors, those providing minimal signal about the treated unit, contribute to the effective saturation parameter  $\kappa$ . Signal concentrates in  $r$  dimensions, leaving a  $(T_0 - r)$ -dimensional noise subspace. Identification diagnostics degrade with  $\kappa$ :

- $\kappa < 1$ : Irrelevant donors span an  $\kappa$ -fraction of the orthogonal subspace, attenuating residuals by  $(1-\kappa)$ . For instance, with  $\kappa = 0.6$ , projection-based diagnostics mask 60% of violations.
- $\kappa \rightarrow 1$ : Detection boundary diverges as  $(1-\kappa)^{-1/4} \rightarrow \infty$ , diagnostic power collapses continuously.
- $\kappa = 1$ : OLS residuals vanish almost surely via random matrix limits, producing perfect fit regardless of  $\delta$ , the population relevance margin. Regularization in dense geometry yields penalty-dependent residuals confounding bias with noise.

This explains why diagnostics fail in large- $N$ , fixed- $T_0$  settings where  $\kappa$  governs feasibility;  $\kappa$  equals the contamination ratio under maximal-rank (unstructured) irrelevant donor contamination, but is smaller in structured cases (e.g., low effective rank  $\text{rank}(X_{\text{irrel}}) < N_{\text{irrel}}$ , yielding  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$ ). The spurious fit trap thus requires both orthogonality (economic irrelevance) and sufficient rank (geometric saturation), presence alone is insufficient if irrelevant donors are low-rank (e.g., highly correlated clusters), preserving diagnostic power.

### Theoretical contributions

First, we prove exact relevance is untestable using projection-based diagnostics (Theorem 4.1): null  $H_0: \delta = 0$  and local alternatives  $H_1: \delta O(T_0^{-1/2})$  have contiguous distributions. No test achieves power exceeding size against such alternatives.

Second, we derive the first sharp detection boundary (Theorem 4.2): violations exceeding

$$\Delta_T = \sigma(1-\kappa)^{-1/4} \sqrt{r/T_0}$$

are detectable with power one, while smaller violations are noise-indistinguishable. This provides explicit thresholds as functions of  $(N, T_0, r, \sigma^2)$ , absent in prior work.

Third, we establish a unified regime taxonomy (Theorems 5.1-5.2):

- (I) Regular (both primitives hold, standard asymptotics),
- (II) Weak conditioning (variance inflation, refinable via extended  $T_0$  or donor pruning removing the colinear ones),
- (III) Structural non-representability (irreducible bias, remediable via donor curation using domain knowledge),
- (IV) Complete breakdown.

This clarifies when refinements help versus when designs fail fundamentally.

### Relation to prior work

These studies provide important insights into specific failure modes, Ferman & Pinto characterize bias geometry, Fernández-Val et al. analyze spurious regressions, Staiger & Stock establish weak-IV asymptotics, but treat them as distinct phenomena.

Our contribution is recognizing these share a common geometric structure governed by the contamination ratio  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$ . By formalizing this unified mechanism, we derive results unavailable in prior work:

- (1) explicit detection boundaries calibrated to contamination levels,
- (2) characterization of the  $\kappa = 1$  phase transition separating partial from complete diagnostic failure,

- (3) impossibility results establishing fundamental testability limits, and
- (4) a regime taxonomy distinguishing refinable from irreducible failures.

This provides both theoretical clarity (understanding when/why methods fail) and practical guidance (calibrating diagnostic power, prioritizing donor curation).

To clarify these increments, Table 1 compares our contributions to key prior works highlighting how we extend geometric insights into explicit, calibratable thresholds and limits, providing a unified diagnostic foundation absent in earlier studies.

**Table 1: Building on Prior Work**

Literature Stream	Representative Papers	What Prior Work Establishes	Unified $\kappa$ -Framework
Geometric Bias in SC	Ferman & Pinto (2021); Botosaru & Ferman (2019)	<ul style="list-style-type: none"> <li>• Bias under donor mismatch has geometric interpretation</li> <li>• Imperfect pre-treatment fit signals potential violations</li> <li>• Simulations show sensitivity to donor pool composition</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Contamination model:</b> Formal <math>\kappa = \text{rank}(X_{\text{irrel}})/T_0</math> parameter</li> <li>• <b>Attenuation mechanism:</b> Prove <math>(1-\kappa)</math> residual masking (Prop 3.1)</li> <li>• <b>Testing theory:</b> Impossibility results (Thm 4.1) and detection boundaries (Thm 4.2)</li> </ul>
		<ul style="list-style-type: none"> <li>• Low-rank projections can fit noise perfectly when <math>N \gg T</math></li> <li>• Asymptotic characterization of spurious fit phenomenon</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Partial regime:</b> Spurious fit begins at <math>\kappa \in (0,1)</math>, not just <math>\kappa &gt; 1</math></li> <li>• <b>Link to relevance:</b> Prove spurious fit masks <math>\delta &gt; 0</math> via geometric saturation</li> <li>• <b>Regularization limits:</b> Show LASSO/Ridge cannot purify contaminated pools</li> </ul>
Spurious Regressions	Fernández-Val et al. (2021)	<ul style="list-style-type: none"> <li>• Consistent estimation under correctly specified low-rank structure</li> <li>• Asymptotic normality and variance decompositions</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Diagnostic foundations:</b> When is factor structure learnable from data?</li> <li>• <b>Relevance vs. conditioning:</b> Separate existence (Def 5.1) from estimability (Def 5.2)</li> <li>• <b>Weak identification:</b> Panel analog of weak IV asymptotics (Thm 4.3)</li> </ul>
Factor Models	Bai (2009); Xu (2017)	<ul style="list-style-type: none"> <li>• Consistent estimation under correctly specified low-rank structure</li> <li>• Asymptotic normality and variance decompositions</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Identification hierarchy:</b> Relevance is logically prior to weighting optimization</li> <li>• <b>Irreducible bias:</b> Prove no weighting scheme overcomes <math>\lambda_0 \notin \text{span}(\Lambda)</math> (Thm 5.1)</li> </ul>
Optimal Weighting	Arkhangelsky et al. (2021); Abadie et al. (2010)	<ul style="list-style-type: none"> <li>• Methods for constructing efficient weights</li> <li>• Pre-treatment fit as informal diagnostic</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Identification hierarchy:</b> Relevance is logically prior to weighting optimization</li> <li>• <b>Irreducible bias:</b> Prove no weighting scheme overcomes <math>\lambda_0 \notin \text{span}(\Lambda)</math> (Thm 5.1)</li> </ul>

Literature Stream	Representative Papers	What Prior Work Establishes	Unified $\kappa$ -Framework
Weak Instruments	Staiger & Stock (1997); Stock & Yogo (2005)		<ul style="list-style-type: none"> <li>• <b>Diagnostic limits:</b> Why fit-based checks fail when <math>\kappa \geq 1</math> (Prop 3.1c)</li> </ul>
		<ul style="list-style-type: none"> <li>• Non-regular asymptotics under weak identification</li> <li>• Variance inflation and non-Gaussian limits</li> <li>• Detection thresholds for first-stage weakness</li> </ul>	<ul style="list-style-type: none"> <li>• <b>Panel translation:</b> Weak conditioning in <math>X'X/T_0</math> mirrors weak instruments</li> <li>• <b>Rate characterization:</b> <math>\lambda_{\min} \asymp T_0^{-\gamma}</math> yields <math>\Omega(T_0^{\gamma-1})</math> variance (Thm 4.3)</li> <li>• <b>Distinction:</b> Temporal vs. cross-sectional conditioning failures</li> </ul>

Synthesis: Our framework unifies these streams by showing geometric bias, spurious fit, and weak identification stem from violations of two primitives, relevance ( $\lambda_0 \in \text{span}(\Lambda)$ ) and conditioning ( $\lambda_{\min}(G_T)$  bounded from zero), with diagnostic degradation governed by contamination  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$ . This yields the first explicit detection boundaries, phase transition characterization, and regime taxonomy for panel causal inference.

## Implications

These limits necessitate:

- (1) prioritizing ex-ante donor curation over ex-post selection,
- (2) enforcing  $\text{rank}(X_{\text{irrel}})/T_0 < 1$  as a hard constraint is sufficient to avoid worst-case diagnostic collapse,
- (3) interpreting diagnostics as one-sided falsification tools,
- (4) recognizing when designs must be abandoned (Regimes III-IV).

## Roadmap

Section 3 characterizes how diagnostics fail: we prove projection residuals attenuate by  $(1-\kappa)$  when irrelevant donors contaminate the pool (Proposition 3.1), establishing the geometric mechanism underlying spurious fits. Section 4 establishes fundamental limits: exact projection-based relevance is untestable even without contamination (Theorem 4.1), and we derive sharp detection boundaries showing projection-based diagnostic power depends on the ratio  $(1-\kappa)^{-1/4}$  (Theorem 4.2). Theorem 4.3 extends these limits to weak conditioning, proving non-regular asymptotics analogous to weak instruments. Section 5 synthesizes results into the identification taxonomy. Section 6 concludes.

## 2. Framework, Identification Primitives and Core Assumptions

This section formalizes the panel data environment and the identification primitives governing counterfactual representability. Unlike standard factor model asymptotics, which typically assume all units' factor loadings are drawn i.i.d. from a common distribution with full support, we explicitly allow for persistent structural heterogeneity in the donor pool. This permits relevance failures to arise endogenously even as the number of donors grows, reflecting empirically common settings where many available controls are economically misaligned with the treated unit.

Our goal is not to construct an adversarial design, but to characterize how approximate irrelevance, donors whose latent structures contain vanishingly small information about the treated unit, interacts with high-dimensional projection geometry to undermine projection-based diagnostic feasibility.

## 2.1 Data Structure and Factor Model

Index the treated unit by  $i = 0$  and donor units by  $i = 1, \dots, N$ . Time is indexed by  $t = 1, \dots, T$ , with treatment occurring after  $T_0$ . Let  $Y_{it}(0)$  and  $Y_{it}(1)$  denote potential outcomes without and with treatment. Only unit 0 is treated post- $T_0$ :

$$Y_{it} = Y_{it}(0) \forall i \neq 0, \forall t, Y_{0t} = Y_{0t}(0) \forall t \leq T_0, Y_{0t} = Y_{0t}(1) \forall t > T_0.$$

Untreated outcomes follow an interactive fixed effects structure:

$$Y_{it}(0) = \lambda_i' f_t + \epsilon_{it},$$

where  $\lambda_i \in \mathbb{R}^r$  is the unit-specific loading vector,  $f_t \in \mathbb{R}^r$  is the common factor vector, and  $\epsilon_{it}$  is mean-zero idiosyncratic noise.

### Pre-treatment Representation:

Define the pre-treatment matrices for the period  $t = 1, \dots, T_0$ :

- **Treated Unit:**  $y_0 = F\lambda_0 + \epsilon_0 \in \mathbb{R}^{T_0}$ ,
- **Donor Pool:**  $X = F\Lambda' + E \in \mathbb{R}^{T_0 \times N}$ ,

where  $F = [f_1, \dots, f_{T_0}]^T$  is the factor matrix,  $\Lambda = [\lambda_1, \dots, \lambda_N]^T$  is the matrix of donor loadings, and  $E$  collects the error terms.

We assume without loss of generality that the data are centered or that the factor structure includes a common trend component. In applied estimators that explicitly include unit intercepts (e.g., Difference-in-Differences or fixed-effects specifications), the projection geometry operates on the subspace orthogonal to the constant vector  $\mathbf{1}_{T_0}$ . This effectively reduces the pre-treatment degrees of freedom from  $T_0$  to  $T_0 - 1$ . Under the asymptotic regime considered here ( $T_0 \rightarrow \infty$ ), this finite-sample adjustment is negligible and does not alter the convergence rates, the definition of the contamination ratio  $\kappa$ , or the impossibility results derived in Section 4.

## 2.2 Identification Primitives

We decompose identification into two geometric primitives: Relevance (Existence) and Conditioning (Estimability).

**Definition 2.1 (Structural Relevance).** A latent counterfactual representation for unit 0 exists if the treated loading lies in the span of the donor loadings:

$$\lambda_0 \in \text{span}(\Lambda)$$

If this holds, there exists a weight vector  $w \in \mathbb{R}^N$  such that  $\lambda_0 = \Lambda^T w$ , implying  $\mathbb{E}[y_0 \mid F] = \mathbb{E}[X \mid F]w$ .

**Definition 2.2 (Conditioning).** The structure is well-conditioned if the Gram matrices of the primitives have minimum eigenvalues bounded away from zero:

- **Cross-Sectional:**  $\lambda_{\min}(\Lambda^T \Lambda / N) > c_\Lambda > 0$  (Donors are not collinear).

- **Temporal:**  $\lambda_{\min}(F^\top F/T_0) > c_F > 0$  (Factors have sufficient variation).

### 2.3 Heterogeneous Donor Alignment and Approximate Irrelevance

Standard factor model asymptotics often assume that all donor loadings  $\lambda_i$  are drawn independently from a common distribution with full support in  $\mathbb{R}^r$ . Under this setup, the relevance condition holds with probability one as the number of donors  $N \rightarrow \infty$ , since the treated unit's loading  $\lambda_0$  is almost surely spanned by the donor loadings. However, this assumption is unrealistic in many empirical applications, where donor pools are assembled based on administrative or institutional boundaries rather than economic alignment, leading to substantial structural dissimilarity between many donors and the treated unit.

To address this, we introduce a framework that accommodates heterogeneity in donor contributions to the treated unit's latent structure. We quantify the alignment of each donor  $i$  with the treated unit using a generic measure  $a_i \in [0,1]$  (for instance, the normalized projection strength onto the treated unit's factors or the correlation between donor and treated trajectories), where higher values reflect strong structural similarity and substantial contribution to representing  $\lambda_0$ , while lower values indicate near-irrelevance with negligible signal. This measure captures not only the presence of shared factors but also their relative magnitudes compared to orthogonal components; for example, if a donor shares partial factors but has dominant orthogonal loadings (e.g., much larger in scale),  $a_i$  can approach zero asymptotically, rendering the donor effectively irrelevant despite nominal overlap.

#### Assumption 2.1 (Approximate Irrelevance with Structured or Unstructured Contamination)

Let the pre-treatment outcome matrix be partitioned as  $X = [X_{\text{rel}}, X_{\text{irrel}}]$ , where relevant donors exhibit alignment bounded away from zero:  $a_i \geq c > 0$  with positive probability. Irrelevant donors satisfy two key conditions that together drive diagnostic degradation:

- (i) **Vanishing Alignment (Economic Irrelevance).** The projection of  $X_{\text{irrel}}$  onto the factor space of the treated unit vanishes asymptotically, ensuring that irrelevant donors provide negligible signal about the counterfactual. This condition holds even for donors with partial factor overlap if orthogonal components dominate the variance in their loadings, as their contribution to the treated unit's structure becomes negligible relative to noise.
- (ii) **Admissible Geometric Complexity (Saturation Mechanism).** The effective rank of  $X_{\text{irrel}}$  may increase with  $T_0$ , yielding the contamination ratio

$$\kappa \equiv \frac{\text{rank}(X_{\text{irrel}})}{T_0},$$

which controls the attenuation in projection-based diagnostics (formalized in Proposition 3.1). This setup encompasses:

- **Unstructured irrelevance:** Fragmented donors span high-dimensional subspaces, with  $\text{rank}(X_{\text{irrel}}) \approx N_{\text{irrel}}$ .
- **Structured irrelevance:** Donors load onto a bounded set of orthogonal factors, resulting in  $\text{rank}(X_{\text{irrel}}) \ll N_{\text{irrel}}$ .

The number of relevant donors satisfies  $N_{\text{rel}} = o(T_0)$ , and idiosyncratic errors follow standard moment conditions. Quantitatively, the aggregate alignment of irrelevant donors satisfies  $\sum_{i \in \text{irrel}} a_i^2 = o(T_0)$ , ensuring their collective signal contribution is asymptotically negligible relative to idiosyncratic variation (e.g.,  $a_i = O(T_0^{-\gamma})$  with  $\gamma > 1/2$  suffices when  $\kappa$  is bounded).

## Interpretation

Irrelevant donors have factor loadings that offer asymptotically vanishing information about the treated unit's counterfactual. As a result, their observed trajectories become dominated by idiosyncratic noise relative to any shared signal. This noise dominance is the fundamental barrier separating relevant from irrelevant donors: even when a donor shares one or more factors with the treated unit, it qualifies as irrelevant if loadings on orthogonal factors are sufficiently large that the signal-to-noise ratio causes  $a_i \rightarrow 0$ . In this regime, the donor's variation becomes indistinguishable from pure noise in the space relevant to the treated unit's counterfactual, rendering it informationally useless despite potential structural overlap.

Economically, irrelevance emerges from both orthogonality and relative strength imbalances. Irrelevant donors might represent:

- Regions shaped by distinct sectoral compositions where shared macroeconomic factors are swamped by sector-specific shocks.
- Institutions operating under different policy regimes where common trends are obscured by regime-specific volatility.
- Units exposed to orthogonal macroeconomic shocks whose magnitude drowns out any weak commonalities.

This approach models smooth, continuous structural heterogeneity in which the transition from relevance to irrelevance is governed by signal degradation rather than abrupt regime separation.

### Definition 2.3 (Structural Donor Types)

We classify the  $N$  donors as follows:

- **Structurally Relevant ( $N_{\text{rel}}$ ):** Donors with alignment  $a_i$  bounded away from zero.
- **Structurally Irrelevant ( $N_{\text{irrel}}$ ):** Donors satisfying the vanishing aggregate condition, yielding negligible signal (strict orthogonality as a special case).

This classification underpins the contamination ratio  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$ , which drives the  $\kappa$ -mechanism explored in Section 3.

## 2.4 Connection to Outcome-Space Geometry

Under the factor model,

$$\mathbb{E}[X_i | F] = F\lambda_i.$$

For irrelevant donors,

$$\|P_F X_i\| / \|X_i\| \rightarrow 0,$$

so their realized outcome vectors behave asymptotically like high-dimensional noise relative to the treated unit's factor space.

Thus, while structurally distinct donors need not be exactly orthogonal, their vanishing alignment implies they increasingly span directions in outcome space uninformative about the treated unit's latent structure.

This approximate irrelevance is sufficient to generate the geometric saturation mechanism studied in Section 3.

## 2.5 Core Statistical Assumptions

We impose the following assumptions to facilitate the high-dimensional asymptotic analysis in Sections 3–5.

**Assumption 2.2 (Strict Latent Exogeneity).** For the pre-treatment period  $t \leq T_0$ , the idiosyncratic errors are strictly exogenous with respect to the factors and treatment assignment:

$$\mathbb{E}[\epsilon_{it} \mid \mathcal{D}, \Lambda, F] = 0$$

and  $\epsilon_{it}$  is independent across units and time.

**Remark:** Strict exogeneity is required because projection estimators utilize the entire pre-treatment block simultaneously. Feedback from  $\epsilon_{it}$  to future factors would invalidate the projection geometry.

**Assumption 2.3 (Low-Rank Structure).** The number of latent factors  $r$  is fixed and finite, with  $r \ll \min(N, T_0)$ .

**Assumption 2.4 (Bounded Moments).** The factors  $f_t$  and loadings  $\lambda_i$  are uniformly bounded. The errors  $\epsilon_{it}$  are sub-Gaussian with variance proxy  $\sigma_\epsilon^2$ .

**Assumption 2.5 (Factor Persistence).** The post-treatment factor path  $\{f_t\}_{t>T_0}$  is generated by the same stochastic process as the pre-treatment path, or satisfies stationarity conditions sufficient to bound the shift in the factor covariance matrix  $\Sigma_F$ .

## 2.6 Summary of the Framework

This framework establishes that identification is not guaranteed by sample size alone.

- Assumption 2.1 (Mismatch) permits the existence of  $\delta > 0$  (Relevance violation) even as  $N \rightarrow \infty$ .
- Definition 2.3 (Irrelevance) formalizes the structurally irrelevant donors in the donor pool.
- Assumption 2.2 (Exogeneity) ensures that any correlation found by the estimator must come from  $\lambda_i$  (signal) or spurious finite-sample noise, not endogeneity.

This sets the stage for Section 3, which proves that when  $\text{rank}(X_{\text{irrel}})$  is large relative to  $T_0$ , projection estimators implicitly weight the irrelevant donors to fit the noise  $\epsilon_0$ , producing spurious fits that mask the relevance violation  $\delta$ . Crucially, this masking is geometric, not statistical; it persists even as sample size grows, representing a fundamental limit of projection-based diagnostics rather than a finite-sample artifact.

## 3. Consequences of Relevance Violations: Motivating Identification Before Estimation

Panel data methods for causal inference: Difference-in-Differences (DiD), Synthetic Controls (SC), and Interactive Fixed Effects (IFE), share a common goal: reconstructing the untreated counterfactual of a treated unit using a linear combination of control units. While these methods differ in how they estimate weights, they all rely on a foundational identification primitive: relevance.

Relevance is the geometric condition that the treated unit's latent structure lies within the span of the donor pool. This condition precedes estimation, inference, and regularization. When it

fails, the counterfactual does not exist in the linear span, and no estimator can recover it within the representational capacity implied by Assumption 2.1.

This section motivates the diagnostic framework by detailing the theoretical consequences of relevance violations. We show that such violations are rarely self-evident: due to high-dimensional noise, non-existent counterfactuals often exhibit excellent pre-treatment fit ("Spurious Fit"), leading to "silent failures" where invalid inferences appear empirically robust.

### 3.1 Defining Relevance and Counterfactual Existence

Relevance is a property of the population factor structure, not the sample realization. We formalize the notion of counterfactual non-existence through the following definition.

#### Definition 3.1 (Outcome-Space Relevance Margin).

Let  $y_0$  and  $X$  denote the pre-treatment outcomes. Conditional on the factor realization  $F \in \mathbb{R}^{T_0 \times r}$ , the squared population relevance margin  $\delta^2$  is the minimum squared projection distance in the pre-treatment outcome space, normalized by time:

$$\delta^2 = \inf_{w \in \mathbb{R}^N} \frac{1}{T_0} \| \mathbb{E}[y_0 | F] - \mathbb{E}[X | F]w \|_2^2.$$

Under the factor model (Assumption 2.3) with  $F'F/T_0 \rightarrow \Sigma_F$ , this outcome-space distance relates to the latent loading mismatch as:

$$\delta^2 = (\lambda_0 - P_\Lambda \lambda_0)' \Sigma_F (\lambda_0 - P_\Lambda \lambda_0) + o(1)$$

where  $P_\Lambda$  is the projection onto  $\text{span}(\Lambda)$  with respect to the inner product induced by  $\Sigma_F$ , where  $\Sigma_F = \mathbb{E}[f_t f_t']$  is the factor covariance. Thus  $\delta = 0$  if and only if  $\lambda_0 \in \text{span}(\Lambda)$ .

*Simplification:* If  $\Sigma_F \approx \sigma_F^2 I$  (spherical factors) or if bounding the error, this simplifies to the Euclidean distance:

$$\delta^2 \asymp \| (I - P_\Lambda) \lambda_0 \|_2^2 \cdot \text{tr}(\Sigma_F)$$

When  $\delta^2 > 0$ , the orthogonal component  $(I - P_\Lambda) \lambda_0$  induces structural bias. Since donor outcomes  $X_t$  depend only on  $\Lambda f_t$ , they contain no information about this component (under Assumption 2.5). Therefore,  $\mathbb{E}[\| \hat{Y}_0(0) - Y_0(0) \|^2]$  is bounded below by  $\| (I - P_\Lambda) \lambda_0 \|^2 \cdot \text{tr}(\Sigma_F)$ , which cannot be eliminated by increasing sample size.

#### Interpretation:

- **Case  $\delta^2 = 0$  (Existence):** The treated unit is generated by the same factors as the donors. A valid counterfactual exists and is uniquely identified (subject to conditioning, see Section 5).
- **Case  $\delta^2 > 0$  (Non-Existence):** The treated unit contains a latent component orthogonal to the donor span. Any linear estimator incurs bias bounded below by  $c \cdot \delta^2$  for some constant  $c > 0$  depending on  $\Sigma_F$ , which cannot be eliminated by increasing sample size.

**Why care?** Ignoring relevance violations leads to profound downstream consequences, as methods proceed under an illusion of identifiability:

- **Biased Point Estimates:** Estimators recover noise-driven approximations rather than true counterfactuals, yielding arbitrary treatment effects.

- **Inference Distortions:** Standard errors and confidence intervals become invalid, as they assume well-conditioned spans. Violations can either inflate variance (when detected) or produce spuriously tight intervals (when masked), fostering overconfidence.
- **Policy Misguidance:** In applications like evaluating public health interventions, spurious results may endorse ineffective policies or dismiss effective ones, with real-world costs (e.g., misallocated resources or delayed reforms).
- **Replication Failures:** Results fail to generalize across donor pools or time periods, undermining scientific credibility.

### 3.2 The Spurious Fit Mechanism: How Irrelevant Donors Mask Identification Failures

In applied practice, the credibility of a counterfactual is often judged by the quality of pre-treatment fit (e.g., pre-treatment RMSE), ideally using holdout validation. The implicit assumption is that low residuals imply  $\delta^2 \approx 0$ .

We show this assumption is false in high-dimensional settings when donor pools are contaminated by irrelevant control units. To formalize this, we first define what constitutes an "irrelevant" donor. To bridge the gap between the economic concept of irrelevance (no shared factors) and the geometric mechanism of spurious fit, we distinguish the structural property from its geometric manifestation.

#### Definition 3.2 (Structural vs. Geometric Irrelevance).

- **Structural Irrelevance:** A donor  $i$  is structurally irrelevant if its alignment measure  $\eta_i$  (as defined in Assumption 2.1) satisfies  $\eta_i \rightarrow 0$ , implying vanishing informational value for the treated unit's latent structure.
- **$\epsilon$ -Geometric Irrelevance:** A donor  $i$  is geometrically  $\epsilon$ -irrelevant if a generic observable alignment measure  $d_i = g(X_i, y_0, F)$  (e.g., pre-treatment correlation, signal-to-noise ratio, or normalized projection strength) satisfies  $d_i < \epsilon$ .

While structural irrelevance is the primitive of interest (capturing continuous heterogeneity in alignment as in Assumption 2.1), diagnostics observe only geometric realizations. The following Lemma establishes that structurally irrelevant donors behave as geometrically irrelevant units with high probability, provided the noise does not overwhelm the signal for relevant units.

#### Lemma 3.1 (Linkage of Structural and Geometric Irrelevance)

Under Assumptions 2.1–2.3:

1. **Irrelevant Donors:** If donor  $i$  is structurally irrelevant ( $\lambda_i = 0$ ), then for any  $\epsilon > O_p(1/\sqrt{T_0})$ , it satisfies geometric irrelevance with probability approaching 1. Its variation is driven entirely by idiosyncratic noise  $\epsilon_{it}$ .
2. **Signal Threshold:** For a structurally relevant donor ( $\lambda_i \neq 0$ ) to be geometrically distinguishable from noise, its signal-to-noise ratio must satisfy  $\|\lambda_i\|/\sigma_\epsilon \gg T_0^{-1/4}$ . Below this threshold, relevant donors are indistinguishable from irrelevant ones.

#### Proof Sketch.

**Part 1:** Sub-Gaussian concentration:  $\|\epsilon_j\|_2 = O_p(\sqrt{T_0}\sigma_\epsilon)$ ;  $\|P_F\epsilon_j\|_2 = O_p(\sigma_\epsilon\sqrt{r})$  (quadratic form, rank  $r$ ). Ratio  $\rightarrow_p 0$  at rate  $1/\sqrt{T_0}$ .

**Part 2:** Signal dominates if  $\sqrt{T_0} \|\lambda_j\| \gg \sigma_\epsilon \sqrt{r}$ , refining to  $T_0^{-1/4}$  via local alternatives/contiguity (Tsybakov bounds).

Full in Appendix A.3.  $\square$

**Remark 3.1 (Interpretation of Irrelevance).**

This definition captures donors whose observed trajectories contain minimal information about the factors driving the treated unit. Mathematically, the condition  $d_i = g(X_i, y_0, F) < \epsilon$  implies that donor  $i$ 's variation is dominated by orthogonal noise or extraneous factors rather than the relevant structural factors  $F$ . Such donors may arise from:

- Units in different economic regimes or institutional contexts
- Geographic regions with fundamentally different structural drivers
- Administrative units included for completeness but lacking causal relevance

In practice,  $\epsilon$  acts as a signal-to-noise threshold. While donors may lie on a continuum of relevance, the asymptotic behavior is governed by those whose signal-to-noise ratio vanishes, which effectively behave as irrelevant in the projection geometry. The key insight is that even donors with small but nonzero projections (e.g., purely spurious correlations where  $\mathbb{E}[P_F X_i] = 0$  but realized projection is nonzero) behave approximately as irrelevant for purposes of identification diagnostics.

We now establish our central result characterizing how irrelevant donors mask relevance violations through geometric saturation of the outcome space.

**Proposition 3.1 (Unified Asymptotic Decomposition Unstructured and Structured Irrelevance).**

Let  $r = y_0 - P_X y_0$  be the projection residual, where  $P_X y_0 = X \hat{w}$  and  $\hat{w} = (X'X)^\dagger X' y_0$  minimize  $\|y_0 - Xw\|_2^2$ . Here  $P_X = X(X'X)^\dagger X'$  denotes the orthogonal projection onto  $\text{span}(X)$ , with  $\dagger$  the Moore–Penrose inverse.

Under Assumptions 2.1–2.5 and Definition 3.2, decompose the donor pool as  $N = N_{\text{rel}} + N_{\text{irrel}}$ . As  $T_0 \rightarrow \infty$  with  $\text{rank}(X_{\text{irrel}})/T_0 \rightarrow \kappa \geq 0$  holding  $N_{\text{rel}}$  fixed or growing sublinearly ( $N_{\text{rel}}/T_0 \rightarrow 0$ ), the normalized squared residual satisfies:

$$\frac{1}{T_0} \|r\|_2^2 \rightarrow_p \underbrace{\mathcal{M}(\delta, \mathcal{U}) \cdot \delta^2}_{\text{Attenuated Relevance Violation}} + \underbrace{(1 - \kappa) \cdot \sigma_\epsilon^2}_{\text{Attenuated Noise}}$$

Where  $\delta^2$  is the squared structural relevance violation (Definition 3.1),  $\sigma_\epsilon^2$  is the idiosyncratic variance,  $\mathcal{M}(\delta, \mathcal{U}) \in [0, 1]$  is the signal masking factor (depends on contamination structure), and  $\kappa$  is the contamination ratio. The noise attenuation  $(1 - \kappa)$  is universal across all cases, while the violation attenuation  $\mathcal{M}(\delta, \mathcal{U})$  exhibits case-specific behavior:

**(i) The Precise Formula (Finite Sample Decomposition):** The effective rank decomposes into signal capture and geometric saturation:

$$\frac{k_{\text{eff}}}{T_0} \approx \underbrace{\frac{r}{T_0}}_{\substack{\text{Classical Term } (\rightarrow 0) \\ \text{Signal + Standard} \\ \text{Overfitting}}} + \underbrace{\frac{\min(\text{rank}(X_{\text{irrel}}), T_0 - r)}{T_0}}_{\substack{\text{Structural Term } (\rightarrow \kappa) \\ \text{Spurious Fit}}}$$

(Note: Relevant donors contribute the factor dimension  $r$ . Irrelevant donors fill the remaining noise space  $(T_0 - r)$  up the rank of their outcome matrix  $X_{\text{irrel}}$ .)

It is crucial to distinguish classical overfitting from the spurious fit mechanism: **Standard Overfitting**: Occurs when relevant donors ( $N_{\text{rel}}$ ) fit the idiosyncratic noise of the treated unit. This is a variance issue ( $r/T_0$  term), remediable by increasing  $T_0$  or regularization; **Spurious Fit ( $\kappa$ -Mechanism)**: Occurs when irrelevant donors fit the noise or the structural mismatch  $\delta$ . This is a bias issue. The projection loads on dimensions orthogonal to the true factors  $F$ , masking the non-existence of the counterfactual. This form of overfitting persists asymptotically if  $\kappa > 0$  and requires donor purification for dense geometries.

**Noise attenuation** is governed by the rank fraction:  $(1 - \kappa) = 1 - \frac{\text{rank}(X_{\text{irrel}})}{T_0}$ .

**Relevance violation attenuation** depends on contamination structure:

- **Case A (Unstructured/Dense Irrelevance)**: When irrelevant donors are mutually uncorrelated with  $\text{rank}(X_{\text{irrel}}) \approx N_{\text{irrel}}$ , the irrelevant subspace  $\mathcal{U}$  is Haar-distributed. By concentration of measure on the Grassmannian:  $\mathcal{M} \rightarrow (1 - \kappa)$  (due to isotropy). Thus, signal and noise are attenuated identically:  $\frac{1}{T_0} \|r\|^2 \rightarrow_p (\delta^2 + \sigma_\epsilon^2)(1 - \kappa)$ .
- **Case B (Structured/Low-Rank Irrelevance)**: When irrelevant donors share orthogonal factors with  $\text{rank}(X_{\text{irrel}}) \ll N_{\text{irrel}}$ , the irrelevant subspace  $\mathcal{U}$  is deterministic. Signal attenuation depends on the principal angle  $\theta$  between  $\delta$  and  $\mathcal{U}$ :  $\mathcal{M} = \| (I - P_{\mathcal{U}})u_\delta \|^2$ . This yields:
  - **Best case (for detection)**:  $\delta \perp \mathcal{U} \Rightarrow \mathcal{M} = 1$  (no relevance violation masking, full diagnostic power)
  - **Worst case (for detection)**:  $\delta \in \mathcal{U} \Rightarrow \mathcal{M} = 0$  (complete relevance violation masking)
  - **Baseline case (isotropic violation under unstructured irrelevance)**: If the violation direction  $u_\delta$  is uniformly distributed over the noise space  $N$  and the irrelevant subspace  $\mathcal{U}$  is Haar-distributed (as in the dense irrelevance regime), then  $\mathbb{E}[\mathcal{M}] = 1 - \frac{\text{rank}(X_{\text{irrel}})}{(T_0 - r)}$ .

In many economic applications, irrelevant donors may be structured (e.g., agricultural states sharing common regional shocks orthogonal to the treated unit). In this case, the effective saturation  $\kappa$  is determined by the rank of the irrelevant block ( $\text{rank}(X_{\text{irrel}}) \ll N_{\text{irrel}}$ ), preserving more projection-based diagnostic signal.

**(ii) The Asymptotic Limit and Structural Saturation**: As  $T_0 \rightarrow \infty$  with  $N_{\text{rel}}$  fixed, the Classical Term is asymptotically negligible ( $r/T_0 \rightarrow 0$ ), standard overfitting becomes negligible, but the Structural Term persists, defining  $\kappa$ , the contamination ratio or the asymptotic structural saturation parameter:

$$\kappa = \lim_{T_0 \rightarrow \infty} \frac{\min(\text{rank}(X_{\text{irrel}}), T_0 - r)}{T_0}$$

The value of  $\kappa$  depends on the correlation structure of the irrelevant donors:

- **Dense irrelevance**:  $\text{rank}(X_{\text{irrel}}) \approx N_{\text{irrel}} \Rightarrow \kappa = \min(N_{\text{irrel}}/T_0, 1)$
- **Structured irrelevance**:  $\text{rank}(X_{\text{irrel}}) = s \Rightarrow \kappa = \min(s/T_0, 1)$

**(iii) Regime Classification**: The behavior of the residual depends on the saturation level  $\kappa$ :

**(a) Well-Specified Regime ( $\kappa = 0$ ):**

$$k_{\text{eff}} \rightarrow r \Rightarrow \frac{1}{T_0} \|r\|^2 \rightarrow_p (\delta^2 + \sigma_\epsilon^2)(1-0)$$

Interpretation: The residual accurately reflects structural bias plus noise. Projection-based pre-treatment fit is a valid diagnostic for relevance violations.

**(b) Partial Spurious Fit ( $0 < \kappa < 1$ ):**

- **Noise:** Always attenuated by factor  $(1-\kappa)$
- **Relevance violation:**
  - Case A Unstructured: Attenuated by  $(1-\kappa)$ , masking is proportional
  - Case B Structured: Attenuated by  $\mathcal{M}(\delta, \mathcal{U}) = \|(I - P_{\mathcal{U}})u_\delta\|^2 \in [0,1]$ , masking varies with alignment

**Interpretation:** Irrelevant donors partially saturate the noise subspace. Detection power degrades, but extent depends on whether contamination is fragmented (Case A) or clustered (Case B).

**(c) Spurious Fit Trap ( $\kappa \rightarrow 1$ ):**

$$k_{\text{eff}} \approx T_0 \Rightarrow \frac{1}{T_0} \|r\|^2 \rightarrow_p 0$$

Interpretation: As the effective rank approaches the full outcome dimension, corresponding to maximal-rank contamination in which irrelevant donors span the orthogonal complement, the donor space saturates the outcome space. Both relevance violation and noise are perfectly interpolated, causing pre-treatment RMSE to converge to zero regardless of the magnitude of the relevance violation  $\delta$ . Consequently, fit-based diagnostics become asymptotically uninformative.

### Remark (Convergence and Practical Applicability)

**(a) Convergence as  $\kappa \rightarrow 1^-$ :** Proposition 3.1 establishes that for any fixed contamination level  $\kappa < 1$ , the residual norm converges to the stated limit as  $T_0 \rightarrow \infty$ . However, we do not claim uniform convergence over all  $\kappa \in [0,1)$  simultaneously. As  $\kappa$  approaches unity, the rate of convergence slows and the approximation quality degrades. This degradation is not a phase transition but a continuous process: diagnostic power erodes smoothly as irrelevant donors increasingly saturate the noise subspace. The limiting behavior as  $\kappa \rightarrow 1^-$  is precisely characterized in Theorem 4.2, which shows that the detection boundary diverges as  $(1-\kappa)^{-1/4}$ , making violations progressively harder to detect even though the residuals remain asymptotically well-defined.

**(b) Finite-Sample Considerations:** The regime classifications in part (iii) rely on asymptotics where the noise subspace dimension  $T_0 - r \rightarrow \infty$ . In empirical applications with short pre-treatment periods, common in comparative case studies (e.g., Abadie et al. 2010 uses  $T_0 = 19$ ), the noise subspace is limited, and high-dimensional random matrix approximations (e.g., Marchenko-Pastur concentration) may exhibit slower convergence. Nonetheless, the finite-sample detection boundary in Theorem 4.2(i-b), which includes an explicit  $\log T_0$  correction, applies exactly for any  $T_0 \geq r$  via sub-Gaussian concentration inequalities. For practical calibration, researchers should interpret regime boundaries as qualitative guides rather than sharp thresholds when  $T_0$  is small.

**(c) Practical Interpretation:** For applied work, the key takeaway is that  $\kappa < 1$  is a necessary condition for diagnostic validity, not a sufficient one. Even when  $\kappa = 0.7$  (seemingly far from saturation), 70% of relevance violations are masked, severely degrading detection power (Theorem 4.2). Conservative practice enforces  $\kappa \leq 0.5$  via ex-ante donor curation, balancing diagnostic sensitivity against sample size. This threshold balances two considerations: (i) preserving at least 50% of diagnostic relevance violation, and (ii) maintaining sufficient sample size for stable weight estimation. Stricter thresholds ( $\kappa \leq 0.3$ ) may be warranted when  $T_0$  is small or violations are expected to be modest.

#### Proof Sketch.

. Full details in Appendix A.3.  $\square$

**Relation to Prior Work:** While existing studies examine these mechanisms separately, our framework unifies them through the evolution of  $k_{\text{eff}}/T_0$ , yielding the sharp threshold  $\kappa = 1$  that separates identifiable from non-identifiable regimes and providing explicit detection boundaries as functions of problem primitives  $(N, T_0, r, \sigma^2)$ .

**Remark 3.2 (Interpretation).** To distinguish the implications for projection-based diagnostics, note that structural irrelevance (vanishing alignment) does not inherently lead to saturation of the outcome space. If irrelevant donors are highly correlated (e.g., a cluster of agricultural states sharing common weather shocks orthogonal to the treated unit's factors), their block has low effective rank, limiting  $\kappa$  and allowing projection-based diagnostics to retain power (as spurious alignments are confined to few dimensions). Conversely, if irrelevant donors are fragmented (e.g., mutually uncorrelated units each driven by independent orthogonal shocks), their effective rank approaches  $N_{\text{irrel}}$ , driving  $\kappa$  toward 1 and enabling the spurious fit trap. Thus, economic irrelevance is necessary but insufficient for complete projection-based diagnostic failure (e.g., the spurious fit trap); it enables power attenuation proportional to  $\kappa$ , but geometric saturation via sufficient rank is required for collapse.

**Table 2: Summary of Unstructured x Structured Irrelevance**

Aspect	Case A: Generic/Dense Irrelevance (Maximal Entropy)	Case B: Low-Complexity Irrelevance (Structured)
Economic Description	Irrelevant donors are fragmented (e.g., independent orthogonal bases; uncorrelated shocks)	Irrelevant donors share few orthogonal factors (e.g., clustered by sector; correlated shocks)
Geometric Effect	$\text{rank}(X_{\text{irrel}}) \approx N_{\text{irrel}}; \kappa \rightarrow 1$ under proportional asymptotics	Effective rank is bounded (e.g., $\text{rank}(X_{\text{irrel}}) \approx \# \text{ clusters} \times \text{factors per cluster}$ ); $\kappa = \text{rank}(X_{\text{irrel}})/T_0 \ll 1$
$\kappa$ -mechanism	$\kappa = 0$ : ideal but difficult to achieve; $0 < \kappa < 1$ : empirically relevant regime with progressive attenuation; $\kappa \rightarrow 1$ : worst-case geometric saturation attainable under fragmented irrelevance.	$\kappa = 0$ : ideal but difficult to achieve; $0 < \kappa < 1$ : empirically relevant regime with limited attenuation; $\kappa \rightarrow 1$ : requires growth in the number of orthogonal factors and is therefore unlikely in low-complexity environments.

Aspect	Case A: Generic/Dense Irrelevance (Maximal Entropy)	Case B: Low-Complexity Irrelevance (Structured)
Diagnostic Power	Degraded or collapsed; spurious fits mask violations (Proposition 3.1(b-c))	Retained; residuals reflect $(1 - \kappa)$ fraction of violations (partial masking limited)
Example	Diverse, uncorrelated irrelevant states each with unique shocks	Midwestern agricultural states sharing crop-yield factors (orthogonal to California's tech factors)
Remedy	Requires donor curation to reduce $\kappa$ ; diagnostics unreliable	Diagnostics could be viable, depending on # clusters and factors per cluster related to $T_0$

**Remark 3.3 (Regularization and the Spurious Fit Trap).** Regularization (Ridge, Elastic Net) modifies the effective degrees of freedom and can, under certain conditions, help mitigate the masking of relevance violations, though it often falls short in dense factor models.

- **Ridge:** In the partial regime ( $0 < \kappa < 1$ ), Ridge reduces attenuation but introduces penalty bias; cross-validation tuning may negate this benefit by driving penalties to zero to maximize fit. However, with theoretically motivated penalties (e.g., fixed  $\lambda \asymp \sqrt{\log N/T_0}$ ), it can stabilize estimates in sparse or low-contamination settings, potentially improving diagnostic sensitivity by shrinking spurious alignments.
- **LASSO/Elastic Net:** While  $\ell_1$  regularization can theoretically "purify" the donor pool if the true counterfactual admits a strict sparse representation (e.g., only neighboring units matter) with sufficient signal strength ( $\min |w_{0i}| \gg \sigma \sqrt{\log N/T_0}$ ) and appropriate tuning, this success is conditional on unverifiable assumptions. In typical dense factor models, LASSO selects donors based on spurious correlation with the noise vector rather than true structural alignment. Because these spurious correlations are mathematically indistinguishable from weak structural signals in the pre-treatment period, LASSO provides no epistemological guarantee of identification, reinforcing the necessity of ex-ante donor curation over algorithmic selection.

See Appendix A.3 for a full theoretical treatment of regularized estimators.

### 3.3 Regularization Does Not Provide Diagnostic Security Against Spurious Fits

While regularization techniques such as LASSO or Elastic Net (Zou & Hastie, 2005) are commonly employed in synthetic control methods to handle large donor pools (e.g., Abadie et al., 2010; Arkhangelsky et al., 2021), they do not offer reliable diagnostic security against the projection-based failures induced by irrelevant donors in dense geometries. These methods solve

$$\hat{w} = \arg \min_w \|y_0 - Xw\|^2 + \lambda(\theta \|w\|_1 + (1 - \theta) \|w\|_2^2),$$

where  $\lambda$  is the penalty parameter and  $\theta$  balances sparsity and shrinkage. The goal is to select a sparse subset of donors while stabilizing weights.

In settings where the true counterfactual admits a sparse representation (e.g., only a small number of economically similar donors matter), regularization can succeed in purifying the donor pool, provided the signal strength is sufficient and penalties are tuned appropriately (e.g., via BIC or theoretical oracle bounds; Bickel et al., 2009). However, in dense factor models, where the treated unit's latent structure is a non-sparse linear combination of many donor loadings, regularization provides no epistemological guarantee of identification. It can produce a stable, sparse estimate that is observationally equivalent to a valid counterfactual, even when the selected donors are chosen based on spurious correlations with the noise vector rather than true structural alignment.

This lack of diagnostic security stems from an omitted variable bias intuition: the structural mismatch  $\delta$  acts as an unmodeled component orthogonal to the relevant donor span. In contaminated pools ( $\kappa > 0$ ), irrelevant donors can spuriously correlate with this mismatch, leading regularization to include them in the active set to minimize the penalized loss. To formalize this, consider the Karush-Kuhn-Tucker (KKT) conditions for LASSO ( $\theta = 1$ ): for an irrelevant donor  $j$ , inclusion in the support requires  $|X_j^\top (y_0 - X\hat{w}_{-j})| \geq \lambda$ , where  $\hat{w}_{-j}$  is the estimate excluding  $j$ . In our setup, irrelevant donors behave noise-like, with  $X_j \approx \epsilon_j$  (idiosyncratic errors), so their correlations with the structural mismatch  $\delta$  or noise  $\epsilon_0$  are random and non-zero.

We derive a condition under which LASSO support recovery fails asymptotically, even when the world might otherwise favor sparsity. Recall the Irrepresentable Condition (Zhao & Yu, 2006; Wainwright, 2009): for exact sign consistency (correctly identifying the true support), irrelevant variables must satisfy

$$\|(X_{\text{rel}}^\top X_{\text{rel}})^{-1} X_{\text{rel}}^\top X_{\text{irrel}}\|_\infty \leq 1 - \eta,$$

for some  $\eta > 0$ , where  $X_{\text{rel}}$  and  $X_{\text{irrel}}$  are the relevant and irrelevant donor matrices. In dense factor models with structured irrelevance (Proposition 3.1 Case B), irrelevant donors share low-rank orthogonal factors, inducing correlations  $X_{\text{rel}}^\top X_{\text{irrel}}/T_0 = O_p(1/\sqrt{T_0})$  if the orthogonal structures are not perfectly decoupled. Using concentration inequalities (Vershynin, 2018, Theorem 4.6.1) on sub-Gaussian matrices, the probability that the Irrepresentable Condition holds decays exponentially as  $\kappa \rightarrow 1$ :

$$P(\text{Irrepresentable Condition holds}) \leq \exp(-c\kappa T_0),$$

for some  $c > 0$ , since the effective dimension for decoupling shrinks to zero, making spurious correlations dominate. The unmodeled  $\delta$  exacerbates this violation by inflating the omitted variable bias in the residuals, forcing the KKT conditions to include irrelevant donors to approximate the missing component. This results in a 'stabilized spurious fit' where the support is non-empty but structurally meaningless, observationally indistinguishable from a valid sparse model.

Even under fixed or theoretically optimal regularization  $\lambda \asymp \sigma\sqrt{\log N/T_0}$  (Bickel et al., 2009), standard oracle inequalities fail because the sparsity assumption is violated by the structural mismatch. Specifically, the unmodeled component  $\delta$  spans a high-dimensional subspace (rank proportional to  $\kappa(T_0 - r)$  as  $\kappa > 0$ ), so the maximum spurious correlation between irrelevant donors and the mismatch scales as  $\max_j |X_j^\top \delta|/\sqrt{T_0} \asymp \sqrt{\log N/T_0}$ . Consequently, whenever the structural violation magnitude satisfies  $\delta^2 \gtrsim \sigma^2 \log N/T_0$ , the spurious gradient contribution

exceeds the noise floor. Since the penalty is calibrated only to suppress idiosyncratic noise  $\sigma\sqrt{\log N/T_0}$ , it is insufficient to suppress the spurious alignment with  $\delta$ .

While cross-validation may drive  $\lambda$  small for prediction, exacerbating overfitting, even optimal  $\lambda$  cannot overcome geometric saturation: LASSO absorbs a  $\kappa$ -fraction of violations into spurious weights, yielding attenuated residuals similar to OLS (Proposition 3.1). Thus, regularization refines Regime II (weak conditioning) but offers no security against Regime III (structural irrelevance), as the resulting fit cannot distinguish spurious from genuine sparsity. Full proofs, including concentration on KKT subgradients, are in Appendix A.3.

To extend the diagnostic limits from Section 4 to penalized estimators, we establish a detection boundary analogous to Theorem 4.2.

**Proposition 3.2 (Detection Boundary for Penalized Estimators)**

Under Assumptions 2.1–2.5 and proportional asymptotics with  $\kappa > 0$ , there exist constants  $c_1, c_2 > 0$  such that:

- (i) Feasibility: If  $\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-3/2} \sqrt{(r + \log N)/T_0}$ , then a test based on the penalized residual norm detects violations with asymptotic power approaching 1.
- (ii) Impossibility: If  $\delta^2 < c_2 \sigma^2 (1 - \kappa)^{-3/2} \sqrt{(r + \log N)/T_0}$ , then no such test achieves power exceeding size +  $o(1)$ .

Moreover, the set of detectable violations is strictly smaller than for OLS (Theorem 4.2), as the  $\log N$  term inflates the boundary by a factor of  $\sqrt{\log N/r}$ , and the contamination exacerbates this via a higher exponent on  $(1 - \kappa)^{-1}$  (specifically,  $-3/2$  vs.  $-1/2$  for OLS on the  $\delta^2$  scale, quantifying the gap as a factor of  $(1 - \kappa)^{-1}$  worse for Lasso under selection uncertainty).

Even if the true counterfactual is sparse (only a few relevant donors exist), a high-dimensional pool of irrelevant donors ( $N_{\text{irrel}} \gg T_0$ ) generates a "shadow" sparse solution composed of irrelevant donors that fits the noise  $\epsilon_0$  better than the true donors fit the signal  $\lambda_0$ . To formalize, consider the LASSO solution path: the probability that a spurious support  $\hat{S}_{\text{spurious}}$  (from irrelevant donors) is selected before the true support  $\hat{S}_{\text{true}}$  is bounded below by

$$P(\text{spurious support selected first}) \geq 1 - \exp(-ck \log N),$$

for some  $c > 0$ , as the maximum spurious correlation scales with  $\sqrt{\log N/T_0}$  (extreme value theory for sub-Gaussians; Vershynin, 2018), dominating the true signal when  $\kappa$  is moderate and  $N$  large. This occurs because, under contamination, the KKT threshold for irrelevant donors is violated with high probability via union bounds over  $N_{\text{irrel}}$ , creating misleading sparse approximations observationally equivalent to genuine ones under local alternatives (via contiguity as in Theorem 4.2).

**Proof Sketch.**

The penalized residual decomposes as  $R = (1 - \kappa)\delta^2 + (1 - \kappa)\sigma^2 + \lambda$  bias term +  $o_p(1)$ . Under sparsity, the active set size is  $O_p(s + \kappa \log N)$ , where  $s$  is true sparsity, yielding variance inflation by  $\log N$ . For the spurious selection bound, apply concentration on  $\max_j |X_j^T(\delta + \epsilon_0)|$  over irrelevant  $j$ , exceeding  $\lambda$  with probability  $1 - \exp(-ck \log N)$  (union bound). This bound holds for local alternatives or signals satisfying an upper bound on true weights (e.g.,  $w_{0i} \lesssim \sqrt{\log N/T_0}$ ), where the true signal strength is on the order of the spurious correlation floor;

stronger signals (e.g., massive loadings) may still be recoverable, but regularization widens the 'Twilight Zone' (Proposition 4.1) by inflating the lower bound of detectable signals with  $\log N$ . Contiguity arguments show indistinguishability for local  $\delta$  scaled by the larger boundary. The  $\log N$  persists even when  $\kappa < 1$  (structured irrelevance): while the irrelevant subspace is low-rank ( $\kappa T_0$ ), the dictionary of  $N$  vectors within it is dense, providing a covering that scales with  $N$  and allowing  $\log N$ -scale extremes in alignment with  $\epsilon_0$ . Full proof in Appendix A.3.  $\square$

To distinguish dense from sparse regimes in practice and assess if LASSO "really worked," practitioners can employ the following fair, empirical checks, balancing optimism for sparse cases with caution for dense ones:

- **Inspect Support Sparsity and Economic Plausibility:** If LASSO selects a small number of donors ( $|\hat{S}| \ll N$ ) that align with domain knowledge (e.g., geographically or economically similar units), this suggests a sparse regime. Conversely, a dense support or inclusion of implausible donors (e.g., unrelated industries) indicates spurious selection in a dense setting.
- **Compare Residuals Across Estimators:** Compute pre-treatment RMSE for LASSO vs. OLS. In sparse regimes, LASSO residuals should be notably smaller without overfitting (e.g., via out-of-sample holdout if  $T_0$  allows splitting). If LASSO residuals are similar to OLS or show instability under bootstrap resampling, suspect a dense regime where regularization fails to purify.
- **Signal Strength Diagnostics:** Estimate the minimum absolute weight in the LASSO support and compare to noise level ( $\min_j |\hat{w}_j| > \hat{\sigma} \sqrt{\log N / T_0}$ , with  $\hat{\sigma}$  from donor residuals). Strong signals (large min) favor sparsity; weak or noisy weights suggest dense contamination.
- **Sensitivity Analysis:** Perturb the pre-treatment data (e.g., add small Gaussian noise  $\sim \mathcal{N}(0, \sigma^2 / T_0)$ ) and re-run LASSO. Stable support across perturbations indicates genuine sparsity; high variability points to spurious correlations in dense models.
- **Cross-Validation vs. Information Criteria:** Use BIC/AIC for tuning instead of CV if suspecting density, these penalize complexity more aggressively. If BIC-selected LASSO yields sparser models with better economic interpretability than CV, it may confirm a workable sparse regime.

These checks promote fairness by acknowledging LASSO's potential in sparse settings while highlighting red flags for dense ones, encouraging practitioners to integrate domain expertise for robust diagnostics.

**Table 2: Sparse vs. Dense Regimes**

Aspect	Sparse Representation	Dense Representation
Weight Distribution	Few non-zero weights (concentrated)	Many non-zero weights (diffuse)
Model Assumption	True model is sparse (e.g., few relevant donors)	Non-sparse, with shared latent factors
Regularization Effectiveness	LASSO purifies pool, recovers true support	LASSO fails due to spurious selections; irrepresentable condition violated

Aspect	Sparse Representation	Dense Representation
Diagnostic Reliability	Pre-treatment fits more trustworthy; less masking	Fits misleading; contamination ( $\kappa$ ) hides biases
Typical Scenarios	Targeted, clustered data (e.g., neighbors only)	Broad factor models (e.g., many overlapping shocks)
Remedies	Algorithmic selection (LASSO) works if tuned properly	Requires ex-ante donor curation; no generic fix
Epistemological Challenge	Verifiable if domain knowledge aligns with support	Hard to distinguish from sparse via data alone; risk of "silent failure"

### 3.4 Implications for Projection-Based Diagnostics

Building on the theoretical decomposition of projection residuals in Proposition 3.1, next we translate the regime-specific behaviors into actionable guidance for empirical researchers. By delineating the practical consequences of varying contamination levels  $\kappa$ , we highlight how diagnostic reliability shifts from robust validation in low-contamination settings to severe masking and outright failure in high-contamination ones, emphasizing the critical role of donor curation in preserving inferential integrity.

**Well-Specified Regime** ( $\kappa = 0$ ): Pre-treatment fit is a valid diagnostic. Detection is feasible when  $\delta^2 \gg \sigma^2/T_0$ , i.e., when signal-to-noise ratio exceeds the factor dimension's effect on degrees of freedom.

**Partial Spurious Fit** ( $\kappa \in (0,1)$ ): Irrelevant donors mask both signal and noise proportionally. Observed residuals underestimate true bias by factor  $(1-\kappa)$ . Detection of relevance violations ( $\delta > 0$ ) requires:

1. **Diagnostic Power Degradation:** Diagnostic power does not vanish effectively at  $\text{rank}(X_{\text{irrel}})/T_0 = 1$ ; it degrades continuously as  $\kappa$  increases. Even moderate contamination ( $\kappa \approx 0.5$ ) forces the residual to understate the true specification error by 50%.
2. **The Necessity of Donor Curation:** Because  $(1-\kappa)$  represents a geometric masking effect, it cannot be "corrected" statistically without knowledge of  $\kappa$  (which is circular). As we prove formally in Corollary 4.1, the gain in diagnostic power from reducing contamination diverges to infinity as  $\kappa \rightarrow 1$ , strictly dominating the variance cost of using a smaller sample size. Therefore, researchers must prioritize ex-ante donor curation (reducing  $N_{\text{irrel}}$  via domain knowledge) over ex-post algorithmic selection in dense geometry.
3. **Conservative Falsification:** Since structured noise would mask less, a common scenario in practice, detected projection-based failures are robust, but "passing" a diagnostic check is a necessary, not sufficient, condition for validity, especially as  $\text{rank}(X_{\text{irrel}})/T_0$  approaches 1.

**Spurious Fit Trap** ( $\kappa > 1$ ): The spurious fit trap establishes that pre-treatment fit quality is fundamentally uninformative about identification when  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$  exceeds unity under saturation of the outcome space. Residuals  $\rightarrow 0$  regardless of  $\delta$ . Standard diagnostics based on projection residuals fail silently, unable to distinguish structural bias from noise.

Projection-based diagnostic Power Degradation:

- (i) Partial Fit, the effective SNR degrades to  $\text{SNR}_{\text{eff}} \approx \frac{\delta^2(1-\kappa)}{\sigma_\epsilon^2}$ . As  $\kappa \rightarrow 1$ , the required violation  $\delta$  diverges.
- (ii) Trap,  $\text{SNR}_{\text{eff}} \rightarrow 0$ .

These results have immediate implications for projection-based diagnostic practice and motivates the sharp detection boundaries we establish in Section 4.

### Donor Purification as Diagnostic Enhancement

Proposition 3.1 suggests a concrete strategy for improving diagnostic power: reduce  $N_{\text{irrel}}$  through ex-ante donor selection in dense geometry. To clarify how partial geometric saturation can arise in empirically relevant settings, consider two scenarios based on the Proposition 99 application in Abadie et al. (2010), which uses 38 U.S. states over a pre-treatment period of length  $T_0 = 19$ .

**Scenario A (Opportunistic Donor Pool with Regularization):** As described in the introduction, suppose the treated unit (California) loads on latent factors associated with technology-intensive industries, environmental regulation, and coastal services, while a subset of donor states is largely orthogonal to these dimensions. For instance, with roughly three dominant factors per group,  $\text{rank}(X_{\text{irrel}}) \approx 9$ , implying an effective saturation index  $\kappa = \text{rank}(X_{\text{irrel}})/T_0 \approx 0.47$ . A researcher applies Elastic Net Synthetic Controls to all 38 states, relying on penalties to select relevant donors. While regularization prevents weight explosion, it fails to distinguish structural alignment from spurious noise correlations in this partially saturated space. The result is a "silent failure": the estimator yields low pre-treatment RMSE and stable, sparse weights, yet effectively interpolates noise rather than recovering a valid counterfactual.

**Scenario B (Curated Donor Pool):** The researcher screens donors using auxiliary information, such as:

- Pre-treatment covariate similarity (economic indicators, demographics)
- Institutional comparability (same regulatory regime, similar governance)
- Geographic proximity or shared economic shocks
- Domain expertise about relevant comparison units

This reduces the pool to 15 donors, with approximately 5 irrelevant units remaining despite screening. If these remaining irrelevant donors are structured, sharing a small number of common orthogonal shocks, the effective rank of the irrelevant block is approximately  $s \approx 3$ , implying an effective saturation index  $\kappa = 3/19 \approx 0.16$ . Our theory implies that in this structured irrelevance case, diagnostic power degrades proportionally to the effective saturation  $\kappa$ : about 16% of the structural violation induced by non-representability is absorbed by spurious geometric alignment, but the approach preserves 84% of the residual signal ( $1 - \kappa$ ) = 0.84, substantially strengthening projection-based fit diagnostics compared to the first case.

**Implication:** The  $\kappa \in (0,1)$  regime implies diagnostic power is not binary but tunable through donor curation. Researchers face a tradeoff:

- **Larger pools** reduce estimation variance (more donors  $\rightarrow$  more information for weight estimation) but degrade diagnostic power (more irrelevant donors spanning different subspaces  $\rightarrow$  more masking of  $\delta$ ).

- **Smaller, curated pools** increase estimation variance (fewer donors → less stable weights) but preserve diagnostic information needed to detect identification failures.

This tradeoff is unavoidable and must be managed explicitly rather than ignored. Our framework suggests:

1. **Prioritize diagnostic power** when  $\text{rank}(X_{\text{irrel}})/T_0$  approaches or exceeds 1.
2. **Accept estimation variance** from smaller donor pools as the cost of maintaining identification credibility.
3. **Use auxiliary information** (covariates, institutional knowledge) to screen donors before any weight estimation.

## From Descriptive Fit to Hypothesis Testing

The spurious fit trap demonstrates that descriptive measures: RMSE,  $R^2$ , residual plots, leave-one-out validation, cannot reliably assess identification in high-dimensional settings. These metrics can approach perfection even when counterfactuals do not exist.

Having established that pre-treatment fit is unreliable when  $\kappa > 0$  and that diagnostic power degrades continuously with  $\kappa$ , we now turn to the fundamental question: What are the theoretical limits of projection-based identification diagnostics? Section 4 proves three core results that establish when violations can and cannot be detected, providing sharp boundaries that delineate the fundamental limits of projection-based identification diagnostics in practice.

## 4. Fundamental Limits of Projection-Based Identification Diagnostics

Throughout this section, we analyze diagnostics based on the projection residual  $r = Y_0 - P_{Y_{\text{don}}} Y_0$ , which underlies standard practices in synthetic control, interactive fixed effects, and generalized DiD. Theorems 4.1–4.2 are method-class-specific: they concern tests/diagnostics that look at fit quality from projections onto donor spaces. The contiguity argument constructs sequences where, under relevance and non-relevance, the joint distribution of projection residuals is asymptotically indistinguishable. This shows that no projection-based diagnostic can separate the two regimes at a rate faster than the detection boundary. However, it does not show that no conceivable statistic using the full joint process of donors and treated could do better. It only demonstrates that, within the projection geometry used by SC/IFE, we hit a fundamental power limit, implying projection-based diagnostics are structurally slower than required for reliable inference, even as data grows.

Section 3 established that when irrelevant donors proliferate, projection methods can achieve nearly perfect pre-treatment fit even in the absence of a valid counterfactual. This phenomenon arises from geometric saturation: as the irrelevant donor pool expands, the column space of  $X$  fills the ambient outcome space, annihilating residual variation. We now investigate the deeper implications of this geometry for projection-based identification diagnostics.

Our focus in this section is on diagnostics based on projection residuals, which underlie standard practices in synthetic control, interactive fixed effects, and related panel methods. We ask: To what extent can the two primitives governing counterfactual existence—relevance and conditioning—be learned from finite pre-treatment data using such diagnostics?

We adopt proportional asymptotics throughout. In particular, the effective rank of the irrelevant donors' matrix grows proportionally with the pre-treatment horizon,

$$\frac{\text{rank}(X_{\text{irrel}})}{T_0} \rightarrow \kappa \in [0, \infty),$$

while the factor dimension remains fixed and small ( $r \ll T_0$ ), and the number of relevant donors  $N_{\text{rel}}$  is fixed or grows sublinearly ( $N_{\text{rel}}/T_0 \rightarrow 0$ ).

This assumption accommodates both unstructured irrelevance, where irrelevant donors are mutually uncorrelated or span independent orthogonal bases (leading to high effective rank approaching  $N_{\text{irrel}}$  and potential saturation  $\kappa \rightarrow 1$ ), and structured irrelevance, where irrelevant donors share a bounded number of orthogonal factors (e.g., clustered by region or sector, keeping  $\text{rank}(X_{\text{irrel}})$  low or slow-growing, thus bounding  $\kappa$  away from 1 even as  $N_{\text{irrel}}$  grows). In the structured case, orthogonality alone does not saturate the noise subspace, preserving diagnostic power against relevance violations; saturation requires sufficient rank in addition to economic irrelevance.

We establish three results. First, exact relevance is untestable using projection-based diagnostics constructed from donor outcome spaces without imposing a minimum separation condition, small violations are fundamentally indistinguishable from noise due to contiguity. Second, large relevance violations admit a sharp detection boundary that diverges as the spurious-fit trap is approached, with explicit rates depending on problem primitives. Third, weak conditioning produces non-regular asymptotics that invalidate conventional inference even when relevance holds. These results reveal intrinsic limits of projection methods and motivate the diagnostic framework developed in subsequent work.

#### 4.1 The Noise Floor: Untestability of Exact Relevance

Relevance requires that the treated unit's factor loading lies exactly in the span of donor loadings,  $\lambda_0 \in \text{span}(\Lambda)$ , or equivalently that the relevance margin  $\delta^2 = 0$ . This is a knife-edge algebraic condition. In finite samples, noise obscures the boundary between in-span and near-span configurations.

The following result formalizes the fundamental indistinguishability for residual-based diagnostics of exact relevance from arbitrarily small violations, establishing a "noise floor" that projection-specific diagnostics cannot penetrate.

##### Theorem 4.1 (Noise Floor for Exact Relevance in Projection-Based Diagnostics)

Under Assumptions 2.1–2.5 and proportional asymptotics with  $N_{\text{rel}}/T_0 \rightarrow \nu \in [0, \infty)$  and  $\kappa \in [0, 1)$  (avoiding the spurious fit trap), consider testing  $H_0: \delta^2 = 0$  versus  $H_{1,T}: \delta_T^2 = c\sigma_\epsilon^2(1 - \kappa)^{-1/2}T_0^{-1/2}$ ,  $c > 0$ .

Without imposing a minimum separation condition on  $\delta^2$ , no projection-based test based solely on pre-treatment outcomes  $\{y_0, X\}$  can simultaneously:

- Control asymptotic size at level  $\alpha$  under  $H_0$ , and
- Achieve power exceeding  $\alpha$  against the local alternatives  $H_{1,T}$  as  $T_0 \rightarrow \infty$ .

The constant  $c$  in the local alternative depends on the factor covariance structure  $\Sigma_F$  through Definition 3.1's metric, but the impossibility result holds for any fixed  $\Sigma_F \succ 0$ .

**Proof Sketch.**

To show impossibility via contiguity, consider the log-likelihood ratio  $\ell_T = \log(p_{H_{1,T}}/p_{H_0})$  for Gaussian residuals in the unattenuated subspace of dimension  $(1 - \kappa)T_0$ . Under  $H_0$ , substitute  $y_0 = F\lambda_0 + \epsilon_0$  to get  $\ell_T = \mu_T' \epsilon_0 / \sigma_\epsilon^2 - \|\mu_T\|^2 / (2\sigma_\epsilon^2)$ , where  $\mu_T$  encodes the local violation with  $\|\mu_T\|^2 / T_0 = \delta_T^2 = c\sigma_\epsilon^2(1 - \kappa)^{-1/2}T_0^{-1/2}$ .

The first term is mean-zero with variance scaling as  $c^2/(1 - \kappa)$ ; the second is  $-c^2/2$ . Thus,  $\ell_T \rightarrow_d N(-c^2/2, c^2/(1 - \kappa))$  under  $H_0$  and symmetric under  $H_{1,T}$ . Positive Hellinger affinity implies mutual contiguity (Le Cam's lemma), bounding power at  $\alpha + o(1)$ . Full details in Appendix A.4.  $\square$

**Remark 4.1 (Interpretation).** This projection-specific impossibility mirrors classical results for weak instruments (Staiger & Stock 1997) and near-unit roots (Elliott et al. 1996). The noise floor  $\sigma_\epsilon^2(1 - \kappa)^{-1/2}T_0^{-1/2}$  for  $\delta^2$  represents the Gaussian detection limit: violations smaller than this are indistinguishable from sampling variation in the factor loadings and idiosyncratic noise. This floor inflates by  $(1 - \kappa)^{-1/2}$  due to contamination, directly linking to the geometric masking in Section 3. As  $\kappa \rightarrow 1$ , the floor diverges, rendering even moderate violations undetectable. This floor holds exactly in finite samples via concentration, with inflation accelerating for  $\kappa$  near 1 even in short  $T_0$  panels, manifesting as steep but continuous power loss rather than abrupt collapse.

## 4.2 A Sharp Detection Boundary for Projection Diagnostics

Although arbitrarily small violations are indistinguishable from exact relevance (Theorem 4.1), larger deviations may be detected. The geometry of Section 3 implies that detectability depends critically on the proportion of irrelevant donors, with power collapsing as  $\kappa \rightarrow 1$ . Let  $r = Y_0 - P_X Y_0$  denote the projection residual. We now establish the feasible detection rate for projection-based diagnostics.

### Theorem 4.2 (Detection Boundary for Projection-Based Diagnostics)

Under Assumptions 2.1–2.5, suppose  $0 < \kappa < 1$ . There exist constants  $c_1, c_2 > 0$ , depending only on distributional parameters and  $r$ , such that:

**(i-a) Asymptotic Feasibility.** If

$$\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r}{T_0}},$$

then there exists a test based on the projection residual norm whose asymptotic power converges to one while controlling size at any fixed level.

**(i-b) Finite-Sample Feasibility.** If

$$\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r + \log T_0}{T_0}},$$

for some constant  $c_1 > 0$  and fixed  $\alpha > 0$ , then violations exceeding this are detectable with probability at least  $1 - \alpha$ .

**(ii) Impossibility.** If

$$\delta^2 < c_2 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r}{T_0}},$$

then no test measurable with respect to the projection residuals attains asymptotic power exceeding size plus  $o(1)$ .

Moreover, as  $\kappa \rightarrow 1$  (saturation of the outcome space), the boundary diverges for fixed  $\delta^2$ , implying that relevance violations become asymptotically undetectable via projection-based residual diagnostics.

**Proof Sketch.**

Decompose  $R = (1 - \kappa)\delta^2 + (1 - \kappa)\sigma^2 + o_p(1)$ . Under  $H_1$ , the mean shifts by  $(1 - \kappa)\delta^2$ .  $\text{Var}(R \mid H_0) \asymp \sigma^2 r (1 - \kappa) / T_0$ , where  $r = \text{rank}_{\text{eff}}((I - P_X)\Sigma_\epsilon)$ .  $\text{SNR} \asymp \delta^2 (1 - \kappa)^{1/2} / [\sigma \sqrt{r/T_0}]$ . For power  $\rightarrow 1$ , require the boundary in (i-a). Finite-sample (i-b) via sub-Gaussian concentration adding  $\log T_0$ . Impossibility via contiguity for local alternatives matching fluctuation order; the experiment reduces to testing the mean of an asymptotically normal statistic with variance  $\asymp \sigma^2 r (1 - \kappa) / T_0$ . Full details in Appendix A.4.  $\square$

**Remark 4.2 (Sharpness of the Rate).** While we have established feasibility of detection at rate  $(1 - \kappa)^{-1/2} T_0^{-1/2}$  and impossibility below this rate up to constants, we conjecture this is minimax optimal (i.e.,  $c_1$  and  $c_2$  can be taken arbitrarily close). Formal minimax lower bounds would require constructing a least favorable prior over violation magnitudes and applying Fano's inequality or Assouad's lemma to bound minimax risk. These refinements are left for future work. For practical purposes, the rate provides actionable guidance on when violations are detectable.

**Remark 4.3 (The Limit of Adaptive Calibration).** A natural methodological response to Theorem 4.2 would be to estimate the saturation parameter  $\kappa$  (e.g., via the eigenvalue spectrum of the donor covariance matrix) and calibrate the detection threshold adaptively using  $(1 - \hat{\kappa})^{-1/2} T_0^{-1/2}$ . However, this adaptive procedure faces a fundamental stability limit. The sensitivity of the detection boundary with respect to  $\kappa$  is governed by the derivative

$$\frac{\partial}{\partial \kappa} [(1 - \kappa)^{-1/2}] = \frac{1}{2} (1 - \kappa)^{-3/2},$$

which diverges to infinity as  $\kappa \rightarrow 1$ . Consequently, in the regime where calibration is most critical (near the spurious fit trap), infinitesimal estimation errors in  $\hat{\kappa}$  translate into unbounded errors in the calculated detection boundary. Since distinguishing weak structural factors from noise eigenvalues is statistically difficult (the "spectral phase transition" problem; see Baik et al., 2005), any estimator  $\hat{\kappa}$  will possess finite variance. This variance is amplified to infinity by the boundary function's singularity, rendering adaptive calibration infeasible. This reinforces the necessity of structural donor purification for dense geometry (ex-ante curation) rather than statistical correction (ex-post adjustment), which would only be effective in sparse geometry using Lasso.

**Remark 4.4 (Robust Variance Estimation via Spectral Decoupling).** To operationalize the test statistic in Theorem 4.2 and avoid the circularity of using attenuated residuals (if  $\hat{\sigma}^2 \propto \|r\|^2$ , the scaling factor  $1 - \kappa$  cancels out), the noise variance  $\sigma_\epsilon^2$  must be estimated from the donor pool alone, independent of the treated unit. Under the factor model (Assumptions 2.1–2.3), we rely

on the spectral properties of the donor covariance matrix  $\hat{\Sigma}_X = X'X/T_0$ . The idiosyncratic variance  $\sigma_\epsilon^2$  is consistently identified by the average of the eigenvalues corresponding to the noise subspace (the "bulk" of the spectrum), utilizing results from random matrix theory (e.g., the Marchenko-Pastur law). Implementation can utilize estimators such as EigenPrism (Janson et al., 2017) or the high-dimensional factor variance estimator of Pelger (2019), which separate latent signal from noise in large panels. Crucially, because this estimator  $\hat{\sigma}_X^2$  depends only on  $X$ , it converges to the true  $\sigma_\epsilon^2$  regardless of the treated unit's relevance violation  $\delta$  or the projection attenuation  $1 - \kappa$ . This ensures the test statistic correctly scales the attenuated residual against the true noise level, preserving the ability to detect deviations.

However, this estimation depends on accurately identifying the number of factors  $\hat{r}$  in the donor pool. If  $\hat{r}$  is underestimated (e.g., failing to detect weak factors), signal eigenvalues leak into the noise bulk, biasing  $\hat{\sigma}^2$  upward. This inflates the detection threshold, rendering the test conservative (reduced power). Conversely, overestimation of  $\hat{r}$  removes large noise components, potentially biasing  $\hat{\sigma}^2$  downward and risking size distortions. Overall, while the theoretical boundary is sharp assuming known parameters, the feasible test is subject to these estimation uncertainties. This exacerbates the circularity noted in Remark 4.3: even if we knew  $\kappa$ , small violations remain undetectable; since we must estimate parameters unstably, practitioners are in a worse position, underscoring the need for ex-ante donor curation.

**Remark 4.5 (Divergence of Detection Boundary).** The minimal detectable violation scales as  $\delta^2 \sim \sigma_\epsilon^2(1 - \kappa)^{-1/2}T_0^{-1/2}$ . As irrelevant donors proliferate ( $\kappa \rightarrow 1$ ), the detection threshold diverges:  $\delta^2 \rightarrow \infty$  even as  $T_0 \rightarrow \infty$ . This formalizes the "continuous power degradation" described in Section 3.4. Example, suppose  $\sigma_\epsilon^2 = 1$ ,  $T_0 = 100$ :

- Well-specified ( $\kappa = 0$ ):  $\delta^2 \sim 0.1$ . Violations  $> 0.1$  are detectable.
- Partial spurious ( $\kappa = 0.5$ ):  $\delta^2 \sim 0.14$ .
- Detection boundary increases by factor  $2^{1/2}$ .
- Near trap ( $\kappa = 0.9$ ):  $\delta^2 \sim 0.32$ . Only large violations detectable.
- Trap ( $\kappa = 1$ ): Detection impossible regardless of  $\delta$ .

**Corollary 4.1 (Dominance of Curation Near Saturation)**

Consider reducing the donor pool by removing  $\Delta$  irrelevant donors, which decreases the effective rank of the irrelevant block by  $\Delta\kappa$ . Define the resulting reduction in the contamination ratio as

$$\eta = \frac{\Delta\kappa}{T_0},$$

where  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$  is the asymptotic structural saturation parameter (Proposition 3.1).

The total change in the effective mean squared error (MSE), where the bias term reflects the squared detection boundary  $\delta_{\min}^2$ , i.e., the maximum undetectable structural violation, decomposes approximately as

$$\Delta\text{MSE} \approx -\frac{\kappa\eta}{(1-\kappa)^{3/2}}T_0^{-1/2} + \frac{\sigma^2\eta T_0}{N^2},$$

where  $k > 0$  is a constant depending on distributional parameters (including  $\sigma^2$ ).

As  $\kappa \rightarrow 1^-$ , the first term (reduction in undetectable bias) diverges as  $(1-\kappa)^{-3/2}$ , while the second term (increase in variance) remains bounded under proportional asymptotics. Thus, near saturation, donor curation yields increasingly large gains in diagnostic power that dominate the modest increase in estimation variance. However, the optimal tradeoff depends on the ratio of relevant donors to pre-treatment periods; when  $N_{\text{rel}}/T_0$  is already small, further reductions may become infeasible.

**Proof.** From Theorem 4.2, the minimum detectable violation scales as

$$\delta_{\min} \sim c\sigma_{\varepsilon}^2(1-\kappa)^{-1/2}T_0^{-1/2}$$

for some constant  $c$  depending on distributional parameters.

This squared detection boundary represents the maximum undetectable structural bias squared. Under standard asymptotics (Regimes I and II), the variance contribution to MSE is of order  $\sigma^2/N$ . Hence, the effective MSE bound takes the form

$$\text{MSE} \approx \kappa(1-\kappa)^{-1/2}T_0^{-1/2} + \frac{\sigma^2}{N}.$$

After removing  $\Delta$  irrelevant donors and thereby reducing the contamination ratio by  $\eta = \kappa - \kappa'$ , we obtain the new parameters

$$\kappa' = \kappa - \eta, N' = N - \Delta = N - \eta T_0.$$

The change in MSE is

$$\Delta\text{MSE} = [k(1-\kappa')^{-1/2}T_0^{-1/2} + \frac{\sigma^2}{N'}] - [k(1-\kappa)^{-1/2}T_0^{-1/2} + \frac{\sigma^2}{N}].$$

For small  $\eta$ , apply a first-order Taylor expansion around  $\eta = 0$ :

$$(1-\kappa')^{-1/2} = (1-\kappa+\eta)^{-1/2} \approx (1-\kappa)^{-1/2} - \frac{1}{2}(1-\kappa)^{-3/2}\eta + O(\eta^2),$$

so the bias term changes by

$$\kappa(1-\kappa')^{-1/2}T_0^{-1/2} - \kappa(1-\kappa)^{-1/2}T_0^{-1/2} \approx -\frac{k\eta}{2}(1-\kappa)^{-3/2}T_0^{-1/2}.$$

The leading-order change in the variance term is

$$\frac{\sigma^2}{N'} - \frac{\sigma^2}{N} \approx \frac{\sigma^2\eta T_0}{N^2}.$$

Combining both contributions and ignoring higher-order terms  $O(\eta^2)$  yields

$$\Delta\text{MSE} \approx -\frac{k\eta}{(1-\kappa)^{3/2}}T_0^{-1/2} + \frac{\sigma^2\eta T_0}{N^2}.$$

As  $\kappa \rightarrow 1^-$ , the first (negative) term diverges in magnitude proportionally to  $(1-\kappa)^{-3/2}\eta T_0^{-1/2}$ , while the second term scales as  $O(\eta)$  and remains bounded under the paper's proportional asymptotics ( $N \asymp T_0$ ). This confirms that near saturation, the diagnostic benefit of curation strictly dominates the variance cost.

However, because  $N = N_{\text{rel}} + N_{\text{irrel}}$  and curation targets only irrelevant donors, the variance penalty  $\sigma^2/N$  assumes that irrelevant donors contribute meaningfully to estimation precision. In cases of strong collinearity or weak conditioning, this cost may be smaller. Moreover, when  $N_{\text{rel}}/T_0$  is already low, aggressive pruning risks reducing the number of relevant donors, potentially making variance inflation non-negligible.  $\square$

**Remark 4.6 (Structured Irrelevance and Targeted Curation).** In structured irrelevance cases (Proposition 3.1 Case B), the effective rank reduction  $\Delta k$  requires targeting donors that span the orthogonal subspace, which may involve removing entire clusters (e.g., all agricultural states sharing common shocks). Random pruning may not suffice; domain knowledge is essential to identify and remove these low-rank blocks, ensuring  $\eta > 0$ .

**Proposition 4.1 (Inference Gap in Projection-Based Diagnostics: The Twilight Zone)**

Under the conditions of Theorem 4.2 with  $\kappa \in [0,1)$ , projection-based diagnostics operate at a fundamentally slower rate than required for valid inference.

Statement. Define detection and inference thresholds:

- Detection:  $\delta_{\text{detect}} = c_1 \sigma_\epsilon (1 - \kappa)^{-1/2} T_0^{-1/4}$  (from Theorem 4.2)
- Inference:  $\delta_{\text{infer}} = c_2 \sigma_\epsilon T_0^{-1/2}$  (for  $\sqrt{T_0}$ -consistent estimation)

Violations in the Twilight Zone  $\mathcal{T} = \{\delta: \delta_{\text{infer}} < \delta < \delta_{\text{detect}}\}$  are:

1. Undetectable: No projection-residual test achieves power  $>$  size  $+ o(1)$
2. Bias-dominant: Asymptotic bias-to-SE ratio  $= \Omega(T_0^{1/4}) \rightarrow \infty$

Rate Gap. The discrepancy arises because:

- Detection requires signal-to-noise  $> 1$  in residuals:  $\delta^2(1 - \kappa) > \sigma_\epsilon^2/\sqrt{T_0}$
- Valid inference requires bias  $\ll$  SE:  $\delta = o(T_0^{-1/2})$

The ratio diverges:  $\delta_{\text{detect}}/\delta_{\text{infer}} = O((1 - \kappa)^{-1/2} T_0^{1/4}) \rightarrow \infty$ .

**Numerical Example** ( $T_0 = 20$ ,  $\sigma_\epsilon = 1$ ,  $r = 2$ ):

$\kappa$	$\delta_{\text{detect}}$	$\delta_{\text{infer}}$	Twilight Zone
0	0.14	0.22	Empty
0.5	0.20	0.22	Narrow
0.8	0.31	0.22	<b>[0.22, 0.31]</b>
0.95	0.63	0.22	<b>[0.22, 0.63]</b>

With  $\kappa=0.8$ , violations producing bias/SE  $\in [1.0, 1.4]$  remain undetectable.

**Implication.** Passing projection-based diagnostics does NOT certify inference validity—only that massive violations are absent. As  $\kappa$  increases, undetectable-yet-bias-dominant violations proliferate. This necessitates ex-ante donor curation (reducing  $\kappa$ ) rather than reliance on ex-post diagnostic "validation."

### 4.3 Weak Conditioning and Non-Regular Inference

We now turn to conditioning. Even when relevance holds ( $\delta^2 = 0$ ), ill-conditioned factor or donor spaces lead to unstable estimation and nonstandard inference. Let

$$\lambda_{\min}(G_T) = \min \left\{ \lambda_{\min} \left( \frac{F^\top F}{T_0} \right), \lambda_{\min} \left( \frac{\Lambda^\top \Lambda}{N} \right) \right\}$$

denote the minimum eigenvalue of the Gram matrices governing temporal (factor) and cross-sectional (donor) conditioning.

#### Theorem 4.3 (Non-Regular Asymptotics under Weak Conditioning)

Suppose relevance holds ( $\lambda_0 \in \text{span}(\Lambda)$ ) and the minimum eigenvalue satisfies  $\lambda_{\min} \rightarrow 0$  at rate  $T_0^{-\gamma}$  for  $\gamma \in (0,1)$ . Consider the least-squares weight estimator  $\hat{w}$ . Then:

**(I) Variance divergence:** The estimation variance satisfies  $\text{Var}(\hat{w}) \sim \Omega(T_0^{\gamma-1})$ , so  $\sqrt{T_0}$ -consistency fails:  $\hat{w} - w = O_p(T_0^{(\gamma-1)/2})$ .

**(II) Rescaled convergence:** The appropriately rescaled estimator  $T_0^{(1-\gamma)/2}(\hat{w} - w)$  converges to a non-degenerate limit whose distribution depends on the eigenvector structure of the weak direction.

**(III) Non-Gaussian limits:** When weak conditioning arises from near-singular donor loadings ( $\lambda_{\min}(\Lambda\Lambda') \rightarrow 0$ ), the asymptotic distribution is non-Gaussian (specifically, a ratio of random variables analogous to weak-IV asymptotics).

**Simplified Gaussian Case.** Under Gaussian factors and errors ( $f_t \sim N(0, I_r)$ ,  $\epsilon_{it} \sim N(0, \sigma^2)$ ), assume the weak conditioning is driven by a single vanishing eigenvalue  $\lambda_{\min}(\hat{G}) \sim cT_0^{-\gamma}$  along eigenvector  $v_N$ . The rescaled error along this direction is

$$T_0^{(1-\gamma)/2}[(\hat{w} - w) \cdot v_N] \rightarrow_d \frac{Z}{\xi},$$

where  $Z \sim N(0, \sigma^2)$  is the projected noise, and  $\xi \sim \chi_1^2/c$  (rescaled chi-squared from the eigenvalue limit). This non-standard ratio invalidates Gaussian inference, with tails heavier than normal. Explicit characterization for general cases requires additional regularity on the factor process and is left to future work. For practical inference, weak-conditioning-robust methods (analogous to Anderson-Rubin in IV) are required.

#### Proof Sketch.

The weight estimator error satisfies  $\hat{w} - w = (X'X)^{-1}X'\epsilon_0$ , where  $X$  is the donor matrix. Under weak conditioning, the smallest eigenvalue of the Gram matrix scales as  $T_0^{-\gamma}$ , causing the inverse to grow at rate  $T_0^\gamma$ . Although the projection of the noise vector  $\epsilon_0$  along this weak direction also has reduced variance, it does not shrink fast enough to counteract the exploding inverse matrix. Specifically, the variance of the estimator along the weak direction scales as  $T_0^{2\gamma-1}$ , which decays strictly slower than the standard  $T_0^{-1}$  rate required for  $\sqrt{T_0}$ -consistency. This "rate deficit" invalidates standard central limit theorems, yielding asymptotic distributions defined by the ratio of the noise projection to the vanishing eigenvalue. In the Gaussian case, the weak direction isolates a single degree of freedom, leading to the chi-squared denominator. Full derivation in Appendix A.4.  $\square$

**Remark 4.6 (Analogy to Weak Instruments).** Theorem 4.3 establishes that weak conditioning in panel data factor models produces asymptotics analogous to weak instruments in IV regression, with the simplified Gaussian limit making the "inference gap" concrete: standard t-tests overreject due to the heavy-tailed ratio distribution.

**Table 4: Analogy to Weak Instruments**

Aspect	Weak Instruments	Weak Conditioning (Panel)
Key Parameter	Concentration parameter $\mu^2 = \pi' Z' Z \pi / \sigma^2$	Minimum eigenvalue $\lambda_{\min}(G_T)$
Non-Regular Regime	When $\mu^2 = O(1)$ : non-regular	When $\lambda_{\min} = T_0^{-\gamma}$ : non-regular
Variance Behavior	2SLS variance diverges	Projection variance diverges
Asymptotic Limits	Non-Gaussian (ratio of normals)	Non-Gaussian (random denominator)
Robust Inference	Anderson-Rubin	Weak-identification robust tests

The key difference: in weak IV, the remedy is often finding better instruments. In weak conditioning, the remedy depends on the source:

- **Temporal weak conditioning** (factors have low variation): extend pre-treatment horizon  $T_0$ .
- **Cross-sectional weak conditioning** (donors are collinear): curate donor pool to reduce redundancy. While neither is universally actionable, this provides more flexibility than the weak IV setting where instrument quality is typically fixed by economic structure.

**Remark 4.7 (Implications for Inference).** Standard inference procedures (asymptotic normality, bootstrap, HAC standard errors) are invalid under weak conditioning because:

- Confidence intervals based on  $\sqrt{T_0}$  have incorrect coverage, often under-covering because the variance divergence is not fully captured by standard covariance estimators.
- Hypothesis tests have incorrect size, rejecting true nulls with probability differing from  $\alpha$ .
- Treatment effect estimates  $\hat{\tau}$  inherit this instability. Developing valid inference under weak conditioning, analogous to weak-IV robust methods (Andrews et al., 2019), is an important direction for future work.

**Remark 4.8 (Contamination and Weak Conditioning).** High contamination ( $\kappa > 0$ ) often induces weak conditioning: irrelevant donors introduce near-orthogonal or collinear directions, making  $\lambda_{\min}(\Lambda^T \Lambda / N) \rightarrow 0$ . The spurious fit trap ( $\kappa \geq 1$ ) typically implies complete breakdown (Regime IV in Section 5), combining non-representability with numerical instability. However, one can have high  $\kappa$  but strong conditioning (e.g., if irrelevants are well-spaced in the noise subspace) or low  $\kappa$  but weak conditioning (e.g., relevant donors with low factor variation). Regime II (Section 5) refers primarily to the latter (classical weak signal from relevants, remediable via extended  $T_0$  or pruning redundants), while Regime III refers to  $\kappa$ -driven structural failures (irreducible bias, remediable only via curation).

#### 4.4 The Spurious Fit Trap and Limits of Projection

The divergence of the detection boundary reflects the geometric mechanism underlying the spurious-fit trap. As the donor space saturates the outcome space ( $\kappa \rightarrow 1$ ), the projection operator eliminates not only noise but also the structural mismatch embodied in  $\delta^2$ . Once  $P_{Y_{\text{don}}}$  approaches the identity, residual-based methods exhaust all degrees of freedom.

Crucially, this failure is not information-theoretic in the full data generating process, but specific to projection-based statistics. The treated unit's outcome  $Y_0$  contains information about  $\delta^2$  in its correlation structure with donors, in higher-order moments, or in auxiliary covariates. Projection-based diagnostics discard this information by construction, focusing solely on the  $\ell_2$  distance  $\|r\|_2^2$ .

Escaping the trap requires abandoning orthogonal projection in favor of alternative geometric diagnostics that:

- Evaluate maximal dependence: Assess the strength of association through non-linear or rank-based metrics that capture fundamental shared dynamics, avoiding the masking effects of conditional linear projection.
- Deploy discriminative screening: Utilize algorithmic frameworks that partition donor pools based on signal-to-noise separation and structural relevance boundaries rather than goodness-of-fit minimization.
- Leverage ensemble aggregation: Exploit recursive or stochastic aggregation mechanisms to identify robust predictive substructures, isolating relevant information from the spurious correlations inherent in high-dimensional noise. Such approaches are developed in our companion empirical work (Part 2). Here we note that the fundamental limits established in Theorems 4.1–4.2 are projection-specific, not universal bounds on identification diagnostics.

#### 4.5 Diagnostics as One-Sided Falsification Tools

Theorems 4.1–4.3 jointly establish an asymmetric role for projection-based identification diagnostics:

- Exact primitives cannot be verified using residual-based fit: Exact relevance ( $\delta^2 = 0$ ) and strong conditioning ( $\lambda_{\min}$  bounded away from zero) are projection-based untestable knife-edge conditions.
- Large violations can be falsified: Violations exceeding problem-specific thresholds ( $\delta^2 > \sigma_\epsilon^2 \sqrt{r/[(1-\kappa)T_0]}$  for relevance,  $T_0^{-\gamma}$  for conditioning with  $\gamma > 0$ ) are detectable with high power.
- Near-boundary cases remain ambiguous: Violations just above or below detection thresholds cannot be reliably classified.

**Implication for practice:** Diagnostics serve as conservative falsification tools, capable of ruling out gross violations but unable to certify that identification holds exactly. This parallels the role of diagnostics in other areas of econometrics:

- Weak instruments: F-statistic  $< 10$  suggests instruments are not grossly weak but doesn't guarantee strong identification (Stock & Yogo 2005).
- Specification tests: Hansen's J-statistic failing to reject does not prove exogeneity, only that overidentifying restrictions aren't grossly violated.

- Balance tests: Covariate balance doesn't prove unconfoundedness, only that observable selection is not extreme.

For relevance:

- Exact span membership ( $\delta^2 = 0$ ) is projection-based untestable.
- Violations smaller than the noise floor are masked by noise.
- When  $\kappa \geq 1$ , all projection diagnostics collapse; even large violations become invisible.

For conditioning:

- Exact non-singularity is projection-based untestable.
- Near-singular designs ( $\lambda_{\min} = T_0^{-\gamma}$  with  $\gamma > 0$ ) produce non-regular behavior that invalidates standard inference even when detected.

This structure is directly analogous to weak instruments, where the first-stage F-statistic detects severe conditioning failures ( $F < 10$ ) but provides ambiguous evidence in borderline cases ( $F \approx 10$ ), much as our detection boundary (Theorem 4.2) separates clearly detectable relevance violations from noise-indistinguishable ones.

#### 4.6 Summary and Forward Outlook

This section establishes sharp limits of projection-based diagnostics:

- **Noise Floor (Theorem 4.1):** Exact projection-specific relevance cannot be tested without minimum separation. Contiguity precludes consistent testing against local alternatives  $\delta^2 \asymp \sigma_\epsilon^2 \sqrt{r/[(1-\kappa)T_0]}$ .
- **Feasible detection (Theorem 4.2):** Violations exceeding  $\sigma_\epsilon^2 \sqrt{r/[(1-\kappa)T_0]}$  are detectable with power approaching one, but the boundary diverges as  $\kappa \rightarrow 1$ , rendering projection-specific diagnostics uninformative in the spurious fit trap.
- **Non-regular asymptotics (Theorem 4.3):** Weak conditioning yields variance divergence and non-Gaussian limits, invalidating conventional inference even when relevance holds.
- **Inference Gap (Proposition 4.1):** Diagnostics are structurally slower ( $T_0^{-1/4}$ ) than required for  $\sqrt{T_0}$ -consistent estimation ( $T_0^{-1/2}$ ), implying undetected violations can dominate bias.

Projection-specific limits: These failures arise from the geometry of high-dimensional orthogonal projection, not from information-theoretic constraints.

Importantly, these limitations are not intrinsic to the identification problem itself, but to the use of projection-based diagnostics. In saturated regimes, orthogonal projections exhaust the ambient space and eliminate the very signal they aim to reveal. Meaningful diagnosis therefore requires approaches that depart from projection geometry.

Having characterized the fundamental limits of projection diagnostics, Section 5 synthesizes these results into a unified identification taxonomy, clarifying when counterfactual recovery is feasible, when it is theoretically possible but practically unstable, and when it faces insurmountable barriers. This taxonomy provides the organizing framework for diagnostic practice and empirical validation in Part 2.

### 5. Identification, Regularity, and Interpretation Under Span-Based Conditions

This section synthesizes the preceding results into a unified characterization of counterfactual existence, point identification, and inferential regularity in span-based panel data methods. We show that causal feasibility is governed by two primitive conditions: **relevance** and **conditioning**, which respectively determine whether a counterfactual representation exists within the donor span and whether its recovery is statistically well-posed. Their interaction yields a complete taxonomy of identification regimes and clarifies the relationship between common panel estimators and the underlying identification environment.

### 5.1 Latent Counterfactual Representability and Relevance

Let untreated outcomes follow the factor structure:

$$Y_{it}(0) = \lambda_i' f_t + \epsilon_{it}$$

and let  $\Lambda = (\lambda_1, \dots, \lambda_N)' \in \mathbb{R}^{N \times r}$  denote the matrix of donor loadings.

#### Definition 5.1 (Latent Counterfactual Representation).

A latent counterfactual representation for treated unit 0 exists if and only if

$$\lambda_0 \in \text{span}(\Lambda)$$

We refer to this condition as relevance.

**Interpretation:** Relevance is a geometric property of the donor pool in latent factor space. It is independent of:

- Estimation procedures (least squares, ridge, synthetic control, etc.)
- Sample size ( $T_0$  or  $N$ )
- Regularization choices (penalty parameters, weight constraints)

When relevance holds, there exists a weight vector  $w$  such that:

$$\mathbb{E}[Y_{0t}(0)] = w^\top \mathbb{E}[X_t(0)].$$

When relevance fails, no estimator can recover the counterfactual within the maintained model class. As established in Section 3, projection-based methods converge to spurious fits that minimize pre-treatment error while retaining non-vanishing post-treatment bias. This yields our first characterization result.

#### Theorem 5.1 (Existence and Point Identification of Counterfactuals)

Under the factor structure (Assumption 2.1) and bounded moments (Assumption 2.3), a latent counterfactual representation exists and is uniquely identified if and only if the relevance condition  $\lambda_0 \in \text{span}(\Lambda)$  holds.

**Moreover:**

**(i) Non-Existence (Irreducible Error):** If relevance fails ( $\delta^2 > 0$ ), then for any measurable estimator  $\hat{Y}_0(0)$  based on donor histories  $Y_{\text{don}}$ , the conditional bias is bounded below:

$$\mathbb{E}[\|\hat{Y}_0(0) - Y_0(0)\|^2 | F] \geq c\delta^2,$$

where  $c > 0$  depends on the factor covariance structure. Consequently, the Mean Squared Error (MSE) is bounded away from zero regardless of sample size.

**Proof Sketch.** Decompose the treated loading into  $\lambda_0 = P_\Lambda \lambda_0 + (I - P_\Lambda) \lambda_0$ , where the orthogonal component induces irreducible error. Full derivation in Appendix A.5.  $\square$

**Remark 5.1 (Uniqueness and Interpretation: Identification vs. Estimation).** When relevance holds, the counterfactual outcome path is uniquely determined (in expectation) by:

$$\mathbb{E}[Y_{0t}(0)] = w^\top \mathbb{E}[X_t(0)],$$

for any weight vector  $w$  satisfying  $\lambda_0 = \Lambda w$ . The Gram may be rank-deficient, making weights non-unique, but the projection is invariant. Theorem 5.1 establishes identification (existence and uniqueness of the counterfactual in the population), not estimation (whether a procedure can recover it with finite sample). Relevance is necessary and sufficient for identification; conditioning (Section 5.2) governs whether stable estimation is possible.

Analogy: In instrumental variables, the rank condition (relevance of instruments) ensures identification of causal effects. Weak instruments (poor conditioning) don't destroy identification but make estimation unstable and inference invalid. Similarly here: relevance ensures counterfactuals exist; conditioning ensures they're stably recoverable.

## 5.2 Conditioning and Inferential Regularity

Existence alone does not guarantee stable recovery. Even when  $\delta^2 = 0$ , the weight vector  $w$  satisfying  $\lambda_0 = \Lambda w$  may be highly sensitive to noise if the donor loadings are nearly collinear or if the factor path exhibits weak variation.

### Definition 5.2 (Conditioning).

The latent structure is **well-conditioned** if the Gram matrices of donor loadings and factor paths have minimum eigenvalues bounded away from zero:

#### (a) Cross-sectional conditioning:

$$\lambda_{\min} \left( \frac{\Lambda' \Lambda}{N} \right) \geq c_\Lambda > 0$$

for some constant  $c_\Lambda$  independent of  $N, T_0$ .

#### (b) Temporal conditioning:

$$\lambda_{\min} \left( \frac{F' F}{T_0} \right) \geq c_F > 0$$

for some constant  $c_F$  independent of  $N, T_0$ .

When conditioning holds in both dimensions, we say the system exhibits persistent conditioning (Assumption 2.5).

### Interpretation:

- **Cross-sectional conditioning** ensures donor loadings span the factor space without redundancy. When  $\lambda_{\min}(\Lambda' \Lambda / N)$  is small, donors are nearly collinear in latent space, many donors provide similar factor exposure, making it difficult to distinguish their contributions. This amplifies estimation error and leads to high-variance counterfactuals.
- **Temporal conditioning** ensures factors exhibit sufficient variation over pre-treatment periods. When  $\lambda_{\min}(F' F / T_0)$  is small, factors are nearly constant or move together,

providing insufficient information to identify their separate loadings. This makes relevance unlearnable from observed data, even when it holds in population.

**Remark 5.2 (Symmetry and Asymmetry in Conditioning).** These two dimensions are symmetric in the factor model representation  $X \approx \Lambda F'$ , since the observed Gram matrix satisfies:

$$\frac{X'X}{T_0} \approx \frac{\Lambda F' F \Lambda'}{T_0} = \Lambda \left( \frac{F'F}{T_0} \right) \Lambda'$$

Ill-conditioning in either  $F'F$  or  $\Lambda' \Lambda$  propagates to  $X'X$ , causing instability in weight estimation.

However, they differ in remedies:

- Poor cross-sectional conditioning  $\rightarrow$  curate donor pool (reduce redundant units)
- Poor temporal conditioning  $\rightarrow$  extend pre-treatment horizon (collect more time periods to capture factor variation).

And in diagnosability:

- Joint conditioning can be assessed directly from the condition number and spectrum of  $X'X$ .
- Isolating the source requires estimating the factor structure (e.g., via PCA), since  $F$  is latent. As a rough diagnostic: if  $\text{cond}(X'X)$  is high but individual donor series  $\{X_i\}$  exhibit substantial temporal variation (high variance), suspect cross-sectional collinearity; if donor variances are uniformly low, suspect weak temporal variation. However, these heuristics are imperfect when the factor structure is complex.

We now formalize how conditioning governs inferential regularity.

**Theorem 5.2 (Conditioning and Regular Inference).**

Suppose relevance holds ( $\delta^2 = 0$ ) so that the latent counterfactual representation is uniquely identified. Consider projection-based estimators obtained as solutions to:

$$\hat{w} = \arg \min_{w \in \mathbb{R}^N} \|y_0 - Xw\|_2^2$$

or regularized variants (ridge, LASSO, elastic net).

**(i) Regular asymptotics:** If persistent conditioning holds (Assumption 2.5), then under standard moment conditions (Assumptions 2.2–2.3), the estimator admits a  $\sqrt{T_0}$ -consistent and asymptotically normal representation:

$$\sqrt{T_0}(\hat{w} - w_0) \xrightarrow{d} N(0, \sigma_\epsilon^2 \Sigma_X^{-1})$$

where  $\Sigma_X = \text{plim}(X'X/T_0)$ , the asymptotic covariance, and inference based on asymptotic normality is valid.

**(ii) Weak identification:** If the minimum eigenvalue converges to zero at rate:

$$\lambda_{\min} \left( \frac{X'X}{T_0} \right) = O(T_0^{-\gamma}) \text{ for some } 0 < \gamma < 1,$$

then no sequence of projection-based estimators achieves uniform  $\sqrt{T_0}$ -consistency. Specifically:

- **Variance Inflation:**  $\text{Var}(\hat{w}) = O(T_0^{-\gamma})$ . While the estimator remains consistent (variance  $\rightarrow 0$ ), it converges strictly slower than the standard  $T_0^{-1}$  rate. The non-Gaussian limiting distribution arises because the projection onto the weak direction has variance scaling with  $\lambda_{\min}$ , while the normalization amplifies this by  $\lambda_{\min}^{-1}$ . This creates a ratio of random quadratic forms similar to weak-IV asymptotics (see Staiger-Stock 1997, Theorem 1). Fully characterizing this distribution is beyond our scope.
- **Rescaled convergence:**  $T_0^{(1-\gamma)/2}(\hat{w} - w_0)$  converges to a non-degenerate limit
- **Non-Gaussian limits:** Asymptotic distributions are generically non-normal, defined by ratios of random variables (noise projected onto weak eigenvectors), invalidating standard  $t$ -tests and confidence intervals.

Standard inference (normal-based confidence intervals,  $t$ -tests) is invalid even when point identification holds.

**(iii) Complete failure:** If  $\lambda_{\min}(X'X/T_0) \rightarrow 0$  at rate  $T_0^{-\gamma}$  with  $\gamma \geq 1$  (e.g., when  $\text{rank}(X_{\text{irrel}})/T_0 \rightarrow \kappa \geq 1$  or factors are constant), then the weights diverge and counterfactual estimates are numerically unstable and uninformative regardless of relevance.

#### Proof Sketch.

**(Part i):** Follows standard OLS asymptotics where the Gram matrix converges to a positive definite limit, preserving  $\sqrt{T_0}$ -consistency via the Central Limit Theorem.

**(Part ii):** The inverse Gram matrix scales as  $\lambda_{\min}^{-1} \asymp T_0^{\gamma}$ . Multiplying this by the noise rate  $O_p(T_0^{-1/2})$  yields an estimation error of order  $O_p(T_0^{\gamma-1/2})$ . The resulting variance  $\Omega(T_0^{2\gamma-1})$  decays strictly slower than the standard  $T_0^{-1}$  rate, creating a "rate deficit" that causes  $\sqrt{T_0}$ -normalized test statistics to diverge.

**(Part iii):** When  $\lambda_{\min} \asymp T_0^{-1}$ , the inverse matrix entries grow linearly with  $T_0$ , causing the weight vector norm to diverge as  $\|\hat{w}\| = O_p(T_0)$ . Full derivation in Appendix A.5.  $\square$

**Remark 5.3 (Why Conditioning Matters Despite Identification).** Theorem 5.2 establishes that relevance and conditioning are logically distinct:

- Relevance determines existence: Does a counterfactual representation exist in the donor span?
- Conditioning determines stability: Can that representation be stably recovered from finite data? Even when  $\delta^2 = 0$  (relevance holds), poor conditioning makes the weight vector  $w$  satisfying  $\lambda_0 = \Lambda w$  extremely sensitive to noise. Small perturbations in observed outcomes lead to large changes in  $\hat{w}$ , producing:
  - Numerically unstable counterfactual estimates
  - Invalid confidence intervals (coverage far from nominal)
  - Spurious treatment effect heterogeneity across time periods This mirrors Weak Instruments in IV: even when instruments are valid (exogenous and relevant in population), weak first-stage F-statistics make 2SLS estimates unstable and inference invalid.

### 5.3 A Unified Taxonomy of Causal Regimes

Combining Theorems 5.1 and 5.2, we obtain a complete partition of the identification space into four exhaustive regimes. Where regimes I–III are separated by identification primitives (relevance and conditioning)

**Regime Classification Table:**

Regime	Relevance	Conditioning	Interpretation	Estimator Behavior	Inference Validity
<b>I: Regular</b>	✓ Holds	✓ Holds	Counterfactual exists and is stably recoverable	$\sqrt{T_0}$ -consistent, asymptotically normal	Valid (standard methods)
<b>II: Weak</b>	✓ Holds	✗ Fails	Counterfactual exists but recovery is unstable	Slower than $\sqrt{T_0}$ , non-Gaussian limits	Invalid (needs robust methods)
<b>III: Structural Non-ID</b>	✗ Fails	✓ Holds	Counterfactual doesn't exist; stable but biased estimates	Converges to biased projection	Invalid ( $L_1$ regularization only if the sparsity condition is met)
<b>IV: Complete Breakdown</b>	✗ Fails	✗ Fails	Counterfactual doesn't exist; estimates unstable	Arbitrary, numerically unstable	Invalid (compound failure)

#### Regime I: Regular Identification

**Conditions:**  $\delta^2 = 0$  and  $\lambda_{\min}(X'X/T_0)$  bounded away from zero.

##### Characteristics:

- Counterfactual representations exist and are uniquely identified
- Weight estimators achieve  $\sqrt{T_0}$ -consistency
- Asymptotic normality holds:  $\sqrt{T_0}(\hat{w} - w_0) \xrightarrow{d} N(0, \sigma_\epsilon^2 \Sigma_X^{-1})$
- Standard inference is valid (Wald tests, normal-based CIs, bootstrap)
- Pre-treatment fit quality provides valid diagnostic information (Proposition 3.1, regime (a))

**Practical implications:** This is the regime implicitly assumed in classical applications of synthetic control (Abadie et al. 2010), interactive fixed effects (Xu 2017), and generalized DiD (Callaway & Sant'Anna 2021). Standard methods succeed without modification.

#### Regime II: Weak Identification (Irregular Asymptotics)

**Conditions:**  $\delta^2 = 0$  but  $\lambda_{\min}(X'X/T_0) = T_0^{-\gamma}$  for some  $\gamma \in (0,1)$ .

##### Characteristics:

- Counterfactual exists and is uniquely identified in population.

- Recovery is unstable: variance inflation occurs (convergence slower than  $T_0^{-1}$ ).
- Asymptotic distributions are non-Gaussian (ratios of quadratic forms).
- Rescaled convergence:  $T_0^{(1-\gamma)/2}(\hat{w} - w_0) \xrightarrow{d}$  non-standard limit.
- Standard inference is invalid: CIs have incorrect coverage, tests have incorrect size.

#### Sources:

- **Temporal:** Weak factor variation over pre-treatment periods (factors nearly constant or perfectly correlated)
- **Cross-sectional:** Near-collinearity among donor loadings (many donors with similar latent positions)

#### Remedies:

- **Extend  $T_0$ :** Collect more pre-treatment periods to strengthen temporal signal (addresses temporal weak conditioning)
- **Curate donors:** Remove redundant units to improve cross-sectional conditioning
- **Regularization:** Ridge regression or Bayesian shrinkage can stabilize point estimates at the cost of introducing bias, but doesn't restore standard asymptotics
- **Robust inference:** Develop weak-conditioning-robust methods analogous to Anderson-Rubin tests for weak IV (future work)

**Key insight:** Regularization (ridge, LASSO) may stabilize point estimation but cannot restore regular asymptotics. Even with penalization, confidence intervals and hypothesis tests remain invalid because the underlying asymptotic distribution is non-Gaussian.

#### Regime III: Structural Non-Identification

**Conditions:**  $\delta^2 > 0$  but weights are statistically stable (e.g., via regularization or well-conditioned donors).

#### Characteristics:

- Counterfactual does not exist within donor pool's representational capacity.
- Estimators converge to best projection:  $\hat{w} \xrightarrow{p} w^* = \arg \min_w \| \mathbb{E}[y_0] - \mathbb{E}[X]w \|^2$ .
- Estimation is numerically stable and achieves  $\sqrt{T_0}$ -consistency for the wrong object.
- Pre-treatment fit can be excellent when  $\kappa \rightarrow 1$  (spurious fit trap, Proposition 3.1 Regime C).
- Post-treatment bias is structural and non-vanishing:  $\mathbb{E}[\hat{Y}_{0t}(0)] - Y_{0t}(0) = \Theta(\delta \cdot \|f_t\|)$ .

**Mechanism:** Irrelevant donors (Definition 3.2) span the residual space orthogonal to true factors, enabling near-perfect interpolation of pre-treatment outcomes while capturing none of the structural signal. This is the silent failure problem: estimates appear credible (low RMSE, stable weights, passes placebo tests) yet are fundamentally invalid.

#### Remedies:

- **No generic statistical remedy exists:** In dense factor environments, the failure is structural; neither sample size accumulation nor standard regularization can resolve non-existence. While  $L_1$  regularization can theoretically isolate a valid counterfactual, this capacity is restricted to regimes satisfying strict sparsity and strong signal conditions (see Appendix A.3); in the general dense case, algorithmic selection fails to distinguish structural alignment from spurious noise correlation.

- **Ex-ante donor curation:** Screen donors using auxiliary information (institutional similarity, economic regime, geographic proximity) to reduce  $N_{\text{irrel}}$  before estimation.
- **Diagnostic vigilance:** Enforce  $\kappa < 1$  as a necessary (not sufficient) condition for diagnostic validity.

#### Regime IV: Complete Breakdown

**Conditions:**  $\delta^2 > 0$  and  $\lambda_{\min}(X'X/T_0) = O(T_0^{-\gamma})$  with  $\gamma \geq 1$ . (or weights explode).

##### Characteristics:

- Counterfactual doesn't exist and recovery is numerically unstable.
- Weight estimates diverge:  $\|\hat{w}\|_2 \rightarrow \infty$  or worse (for unregularized estimators).
- Prediction variance explodes even in-sample.
- Both point estimates and inference are invalid.
- Diagnostics fail: residuals  $\rightarrow 0$  (masking non-existence) while numerical instability produces high variance.

**Mechanism:** Compound failure where irrelevant donors create spurious dimensions while near-singular design amplifies noise. The worst of both worlds. As established in Theorem 4.1, the fundamental "Noise Floor" for detecting relevance violations scales as  $T_0^{-1/2}$ . In Regime IV, weak conditioning inflates this floor by the factor  $T_0^{\gamma/2}$  (Theorem 4.3), effectively widening the "Twilight Zone" (Proposition 4.1) until even massive structural violations become statistically undetectable amidst the variance explosion.

**Remedies:** Same as Regime III (donor curation to restore relevance) plus same as Regime II (donor curation to remove redundant units and extending  $T_0$ , or regularization to improve conditioning). However, neither alone is sufficient, both failures must be addressed simultaneously.

**Corollary 5.1 (Asymptotic Signatures of Regimes II and III).** Under the conditions of Theorems 5.1 and 5.2:

- **Regime II (Weak Conditioning,  $\delta^2 = 0$ ,  $\lambda_{\min}(G_T) = O(T_0^{-\gamma})$ ):** The weight estimator  $\hat{w}$  exhibits variance explosion:  $\|\hat{w}\|_2$  diverges or oscillates as  $T_0 \rightarrow \infty$  (or regularization penalty  $\rightarrow 0$ ). Counterfactual estimates are unstable, with bimodal or heavy-tailed distributions.
- **Regime III (Structural Non-ID,  $\delta^2 > 0$ , Persistent Conditioning):**  $\hat{w} \rightarrow_p w^*$  (a stable, finite limit, the biased projection onto the deficient span). Counterfactual estimates converge stably to the wrong value, often with spuriously tight confidence intervals.

**Proof Sketch.** In Regime II, the inverse Gram amplifies noise without bound (Theorem 4.3). In Regime III, the projection minimizes MSE onto the wrong span, converging stably by persistent conditioning. Full in Appendix A.5.  $\square$

**Interpretation:** These signatures operationalize the taxonomy: oscillating weights signal Regime II (remediable via extended  $T_0$  or pruning redundants), while stable weights with "too good" fit (low residuals despite domain mismatch) signal Regime III (remediable via curation to restore span). Compound failures (Regime IV) are multiplicatively worse: weak conditioning inflates the noise floor from Section 4 by  $T_0^{-\gamma/2}$ , making relevance violations even harder to detect.

**Remark 5.4 (Empirical Classification):** While the regime taxonomy is defined via population primitives ( $\delta^2$ ,  $\lambda_{\min}$ ), practitioners face the challenge of assessing which regime applies to their data. Exact classification is impossible (Theorem 4.1), but researchers can:

- Compute condition numbers of  $X'X/T_0$  to detect weak conditioning,
- Enforce  $\kappa < 1$  to preserve diagnostic validity (Proposition 3.1),
- Use the detection boundary (Theorem 4.2) to bound minimum testable violations.

Part 2 develops practical diagnostic procedures that operationalize these principles without requiring knowledge of  $r$ ,  $\hat{w}$ ,  $\delta^2$  or  $\lambda_{\min}$ .

## 5.4 Interpretation and Relation to Existing Methods

The relevance-conditioning framework clarifies the behavior of widely used panel estimators and the role of methodological refinements.

**Difference-in-Differences (DiD):** Standard DiD assumes parallel trends, a degenerate special case of our factor model where loadings on time-varying factors are homogeneous:  $\lambda_i = \lambda$  for all  $i$ . The estimator is:

$$\hat{\tau}_{\text{did}} = (\bar{Y}_{1,\text{post}} - \bar{Y}_{1,\text{pre}}) - (\bar{Y}_{0,\text{post}} - \bar{Y}_{0,\text{pre}}),$$

where subscripts denote treated (1) vs. control (0) groups, and pre/post periods.

Connection to the framework:

- **Implicit relevance:** DiD assumes all units share the same loading on a single time trend (1-factor model). Relevance holds trivially under homogeneous trends but fails if the treated unit has unique loadings ( $\delta^2 > 0$ ).
- **Conditioning via pooling:** DiD pools controls into a single average using fixed weights  $1/N$ , improving stability in Regime II (weak conditioning) when  $N \gg T_0$ , but it assumes no heterogeneity, making it vulnerable to confounding.
- **Failure mode under contamination:** Violations of parallel trends correspond to relevance failures in a 1-factor model. In staggered adoption designs with many controls, irrelevant units introduce spurious alignments, masking violations via the  $\kappa$ -mechanism (Proposition 3.1), producing excellent event-study balance (pre-treatment zeros) while retaining biased treatment effects (Regime III). Generalized DiD (Callaway & Sant'Anna, 2021; Sun & Abraham, 2021) allows heterogeneous trends, implicitly expanding the factor space and addressing some Regime II issues, but still fails under structural non-representability without curation.
- **Implication:** Pre-treatment diagnostics (e.g., event-study plots) can be misleading in contaminated pools ( $\kappa > 0$ ), as irrelevant controls absorb violations. Discrepancies between DiD and donor-based methods (e.g., SC/IFE) on curated subsets signal Regime III failures, where bias is irreducible, prioritizing ex-ante donor curation for robust identification over generalized extensions.

**Synthetic Control (SC):** Standard synthetic control methods (Abadie et al., 2010) construct weights  $\hat{w}$  to approximate the treated unit's pre-treatment outcomes:

$$\hat{w} = \arg \min_w \|y_0 - Xw\|^2 \text{ s.t. } w \geq 0, \sum w_i = 1.$$

Connection to the framework:

- **Implicit relevance:** SC assumes the treated unit's loading  $\lambda_0$  lies in the donor linear span. The convex hull restriction ( $w \geq 0, \sum w_i = 1$ ) is strictly stronger than mere span inclusion; a unit may lie in span ( $\Lambda$ ) but outside the positive convex hull of donor loadings.
- **Conditioning via constraints:** The simplex constraint stabilizes weights in Regime II (weak conditioning) by preventing extrapolation, but it raises the threshold for perfect spurious fits relative to unconstrained projections, substituting linear rank saturation ( $\kappa \rightarrow 1$ ) with slower logarithmic saturation governed by extreme value theory of the convex hull, scaling as  $(2\sigma^2 \log N_{\text{irrel}})/T_0$  instead of  $N_{\text{irrel}}/T_0$ . Nonetheless, moderate violations may still be partially masked if  $\delta^2$  aligns favorably within the positive hull. This scenario is realistic in practice, since panel data with short to medium pre-treatment periods and a substantial number of donors often result in rank ( $X_{\text{irrel}}$ ) approaching  $T_0$ , especially when irrelevant donors span multiple orthogonal bases. These dynamics are formalized in Proposition 5.3 in Appendix A.5.
- **Failure mode under contamination:** When relevance fails ( $\delta^2 > 0$ ), SC converges to a spurious fit that minimizes pre-treatment error (often hitting the simplex boundary) while retaining irreducible structural post-treatment bias (Regime III). Irrelevant donors enable partial absorption of violations via the  $\kappa$ -mechanism (Proposition 3.1), with impossibility results applying a fortiori (Theorem 4.1).
- **Implication:** Pre-treatment fits can be misleading in contaminated pools ( $\kappa > 0$ ), as the convex constraint attenuates but does not eliminate masking. Discrepancies between SC and unconstrained methods (e.g., IFE) on curated donors signal Regime III failures, emphasizing ex-ante donor curation to avoid nonexistent counterfactuals.

**Regularization (Ridge, LASSO, Elastic Net):** Penalized estimators solve:

$$\hat{w} = \arg \min_w \|y_0 - Xw\|^2 + \lambda \|w\|_p,$$

where  $p = 2$  for Ridge,  $p = 1$  for LASSO, or a mixture for Elastic Net ( $p = \theta \cdot 1 + (1 - \theta) \cdot 2$ ),  $\lambda > 0$  is the penalty, and tuning balances fit and shrinkage.

Connection to the framework:

- **Implicit relevance:** Regularization assumes the donor span can approximate the treated unit; it does not enforce or test relevance directly. If  $\delta^2 > 0$ , shrinkage operates on the deficient span, yielding a biased solution without resolving structural non-representability (Regime III).
- **Conditioning via penalties:** Penalties stabilize weights when the Gram matrix  $X^\top X$  is nearly singular, addressing Regime II (weak conditioning) by trading bias for variance reduction, analogous to weak instruments (Staiger & Stock, 1997).
- **Failure mode under contamination:** Irrelevant donors amplify spurious correlations; in dense geometries ( $\kappa > 0$ ), regularization absorbs a  $\kappa$ -fraction of violations into shrunken weights, attenuating residuals via the  $\kappa$ -mechanism (Proposition 3.1). LASSO may select irrelevant donors if the irrepresentable condition fails, while Ridge/Elastic Net yields dense, biased fits without purifying the pool.
- **Implication:** Regularization improves MSE in well-identified but ill-conditioned settings (Regime II), but offers no diagnostic security against fundamental failures (Regimes III–IV), pre-treatment fits remain misleading, and inference is invalid without regular asymptotics (Theorem 5.2). Discrepancies between regularized and unregularized fits on curated pools signal irreducible bias, prioritizing ex-ante donor curation over penalties.

**Interactive Fixed Effects (IFE):** IFE methods (Bai, 2009; Xu, 2017) jointly estimate latent factors and loadings across all units (including the treated):

$$(\hat{F}, \hat{\Lambda}) = \arg \min_{F, \Lambda} \|Y - F\Lambda^\top\|^2 + \text{penalties},$$

where  $Y$  is the full  $N \times T_0$  pre-treatment matrix,  $F$  is  $T_0 \times r$ ,  $\Lambda$  is  $N \times r$ , and penalties (if any) control rank or overidentification.

Connection to the framework:

- **Implicit relevance:** IFE assumes that all units, including the treated, share the same low-rank factor structure. When the model is correctly specified with sufficient rank  $r$ , relevance holds by construction, as  $\lambda_0$  is estimated directly within the common span.
- **Conditioning via pooling:** Joint estimation leverages all  $NT_0$  observations, dramatically improving factor identification and eigenvalue conditioning compared to SC (which uses only  $T_0$  equations for the treated unit). This makes IFE far more robust to Regime II (weak conditioning) than donor-based methods.
- **Failure mode under contamination:** If the treated unit has unique factors not shared by donors ( $\delta^2 > 0$ ), forcing it into the common low-rank span induces classic omitted-variable bias. Irrelevant donors contaminate the pooled estimation, biasing the global factors toward spurious dimensions that better fit the entire panel, thereby masking the treated unit's relevance violation through the same  $\kappa$ -mechanism (Proposition 3.1, structured case).
- **Implication:** Pre-treatment fit in IFE can be even more misleading than in SC, achieving near-perfect fits by over-absorbing idiosyncratic shocks or structural mismatches into the common factors. Large discrepancies between SC (unit-specific convex weights) and IFE (pooled factors) on the same donor pool are a powerful diagnostic signal of Regime III (structural non-representability), as they reveal that the treated unit cannot be faithfully represented either way, strongly motivating ex-ante donor curation before applying any method.

**Augmented Synthetic Control (ASC):** ASC methods (Ben-Michael et al., 2021) combine outcome regression with synthetic control:

$$\hat{Y}_{0t}(0) = \hat{\mu}(X_{0,\text{pre}}, t) + \hat{w}^\top (X_t - \hat{\mu}(X_{\cdot,\text{pre}}, t)),$$

where  $\hat{\mu}$  is a flexible regression function (e.g., Ridge, Random Forest) fitted on pre-treatment data, and  $\hat{w}$  are synthetic weights.

Connection to the framework:

- **Implicit relevance:** ASC assumes the donor span can approximate the treated unit's structure after adjusting for observable covariates via  $\hat{\mu}$ . Relevance still requires  $\lambda_0$  to lie in the augmented span (donors plus covariates); if  $\delta^2 > 0$ , the orthogonal component  $\lambda_0^\perp$  induces irreducible bias.
- **Conditioning via regularization:** Regularized regression in  $\hat{\mu}$  improves stability under weak conditioning (Regime II), trading bias for variance, while SC weights handle relevance.
- **Failure mode under contamination:** Irrelevant donors contaminate the span, allowing  $\hat{\mu}$  to exploit spurious correlations with noise  $\epsilon_0$  or mismatch  $\delta^2$ , attenuating residuals via the  $\kappa$ -mechanism (Proposition 3.1). Internal augmentation (e.g., lags from deficient  $\Lambda$ ) fails; external covariates must span the gap to resolve Regime III.

- **Implication:** Augmentation reduces variance but is no free lunch for identification, pre-treatment fits can mislead in contaminated pools ( $\kappa > 0$ ) unless external data fixes relevance. Discrepancies with standard SC on curated pools signal irreducible bias (Regime III), prioritizing ex-ante curation and external augmentation.

**Matrix Completion:** MC methods (Athey et al., 2021) imputes the counterfactual by completing the low-rank matrix of untreated outcomes via nuclear-norm regularization:

$$\hat{L} = \arg \min_L \|L\|_* + \frac{\lambda}{2} \sum_{(i,t) \in \Omega} (Y_{it} - L_{it})^2$$

where  $\|L\|_*$  is the nuclear norm (sum of singular values),  $\Omega$  is the set of observed (untreated) entries, and  $\lambda > 0$  controls the rank-fit trade-off. Post-treatment counterfactuals are the imputed entries for the treated unit.

Connection to the framework:

- **Implicit relevance:** MC assumes the full untreated outcome matrix is approximately low-rank across all units and time. Relevance holds if the treated unit's latent loading  $\lambda_0$  lies in the column space of this global low-rank approximation.
- **Conditioning via regularization:** The nuclear-norm penalty enforces low rank (analogous to interactive fixed effects), pooling information across the entire panel to stabilize recovery even when  $N \gg T_0$ , without requiring explicit choice of factor dimension  $r$ .
- **Failure mode under contamination:** Irrelevant donors introduce spurious dimensions into the observed matrix; the low-rank completion can exploit these to fit noise  $\epsilon_0$  or the structural mismatch  $\delta^2$ , attenuating residuals via the  $\kappa$ -mechanism (Proposition 3.1).
- **Implication:** Pre-treatment fit can be misleading in dense geometries ( $\kappa > 0$ ). Large discrepancies between MC and SC (or IFE) on a curated donor pool signal potential Regime III (structural non-representability), where bias is irreducible and cannot be resolved by regularization.

**Synthetic Difference-in-Differences (SDID):** SDID methods (Arkhangelsky et al., 2021) combine synthetic controls with difference-in-differences by estimating unit weights ( $\hat{\omega}$ ) and time weights ( $\hat{\lambda}$ ) via penalized optimization, then fitting a weighted two-way fixed effects regression including unit intercepts ( $\alpha_i$ ) and time effects ( $\beta_t$ ):

$$\hat{\tau}_{\text{sdid}} = \arg \min_{\tau, \mu, \alpha, \beta} \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \hat{\omega}_i \hat{\lambda}_t,$$

where  $W_{it}$  is the treatment indicator, and the weights are chosen to balance pre-treatment outcomes.

Connection to the framework:

- **Implicit relevance:** SDID assumes a low-rank factor structure similar to ours, with relevance required for unbiased estimation. The unit intercepts absorb additive heterogeneity, equivalent to demeaning the data across time for each unit, projecting onto a subspace orthogonal to the constant vector. This reduces the effective outcome dimension

by approximately 1 per fixed effect type (unit and time), consuming degrees of freedom and making the saturation threshold ( $\kappa = 1$ ) slightly stricter (effective  $T_0$  smaller). However, in the paper's asymptotics ( $T_0 \rightarrow \infty$ , fixed  $r$ ), this adjustment is negligible, as the noise subspace remains  $\approx T_0 - r$ .

- **Conditioning via regularization:** Time weights modify the inner product in the Gram matrix ( $X^T \Omega X$ ,  $\Omega = \text{diag}(\hat{\lambda})$ ), potentially improving temporal conditioning (Definition 2.2(b)) but not altering relevance or the  $\kappa$ -driven saturation mechanism (Proposition 3.1). SDID thus addresses Regime II (weak conditioning) via weighting, but like SC, fails in Regime III (structural non-ID) without donor curation.
- **Failure mode under contamination:** If relevance fails, SDID converges to a biased weighted projection, with pre-treatment fit masked similarly to SC. Irrelevant donors introduce spurious alignments, allowing the low-rank approximation to absorb a  $\kappa$ -fraction of violations into the weighted fits.
- **Implication:** Pre-treatment diagnostics can be misleading in contaminated pools ( $\kappa > 0$ ). Discrepancies between SDID and standard SC/IFE on curated donors signal Regime III failures, emphasizing the need for ex-ante curation over algorithmic adjustments.

**Weak Identification Analogy:** The conditioning failures in Regime II bear a close analogy to weak instruments in instrumental variables (IV) regression (Staiger & Stock, 1997):

- **Key Similarity:** In both cases, point identification holds in population (IV exclusion restriction is satisfied; panel relevance is satisfied), but finite-sample estimation is unstable, leading to variance inflation and invalid inference under standard asymptotics (Theorem 4.3).
- **Key Difference:** Unlike weak IV, where instrument quality is often fixed by economic structure (e.g., no feasible alternatives), weak conditioning in panel methods offers two remedial margins: (i) temporal, by extending  $T_0$  if historical data are available to strengthen factor Gram matrices; and (ii) cross-sectional, by curating donors to prune irrelevant or collinear units. While neither margin is universally actionable (e.g., data constraints may limit extensions), this duality provides greater flexibility than typical IV settings, potentially shifting from Regime II to Regime I via refinements (Theorem 5.2).

## 5.5 Summary

Key takeaways:

1. Structural non-identification (Regime III) arises when relevance fails, regardless of sample size or estimator choice. More data or better algorithms cannot solve non-existence.
2. Irregular inference (Regime II) arises when conditioning fails, even when identification holds. Standard asymptotic theory breaks down, requiring weak-conditioning-robust methods.
3. The spurious fit trap ( $\kappa \geq 1$ ) masks relevance violations through geometric saturation. Pre-treatment fit becomes uninformative about identification.
4. Methodological refinements (augmentation, regularization, optimal weighting) address estimation variance but do not overcome identification failures unless external data spans the gap. They improve Regime II, cannot fix Regime III.

This framework provides a unified explanation for the successes and failures of common panel data methods and delineates the fundamental limits of span-based counterfactual estimation. Sections 3–5 together establish both the theoretical boundaries of what can be identified and diagnosed, and the practical guidance for empirical implementation.

## 6. Conclusion

This paper develops a unified identification framework for panel data methods in causal inference, emphasizing that counterfactual validity hinges on two geometric primitives: relevance (the treated unit's latent factors residing in the donor span) and conditioning (stable recovery of this representation). We demonstrate that projection-based diagnostics degrade systematically due to the effective-irrelevant-donor-rank-to-time ratio  $\kappa$ , distinct from the overall donor-to-time ratio  $J/T_0$ .

Projection-based diagnostic power erodes immediately for  $\kappa > 0$ , worsening progressively toward unity. In partial spurious fit regimes ( $0 < \kappa < 1$ ), irrelevant donors mask a  $\kappa$ -fraction of relevance violations via noise subspace saturation; e.g.,  $\kappa = 0.6$  conceals 60% of structural bias, making moderate violations undetectable even if  $J < T_0$ . This is our core insight: contamination proportionally impairs diagnostics prior to total failure at  $\kappa = 1$ .

Regularization (e.g., Synthetic Controls, Elastic Net) does not mitigate this degradation. While effective for variance control (conditioning), it cannot separate true signals from spurious alignments (relevance). For  $\kappa > 0$ , sparsity penalties select irrelevant donors via correlations with noise  $\epsilon_0$ , yielding stabilized spurious fits that shrink degrees of freedom without pool purification. Residuals mix penalty bias with structural mismatch, undermining isolation of violations.

We outline projection-based diagnostics' limits: exact relevance is untestable due to contiguity of in-span and near-span distributions. A sharp detection boundary scales with  $\kappa$ , where rising  $\kappa$  inflates required violation magnitudes for detection. Thus, diagnostics serve as one-sided falsifiers, ruling out major breaks for low  $\kappa$ , but impotent amid irrelevant dilution.

### Implications for Research Practice

Theoretical constraints demand empirical shifts:

- **Donor Curation Priority:** Avoid "kitchen sink" inclusion, which erodes power. Use ex-ante domain knowledge screening to minimize  $\kappa$ , outperforming ex-post algorithms.
- **Diagnostic Caution:** Low pre-treatment RMSE in  $\kappa > 0$  only excludes extreme breaks, not confirming counterfactual validity.
- **$\kappa < 1$  as Constraint:** Treat  $\kappa < 1$  as essential for avoiding diagnostic collapse, targeting lower values for partial contamination resilience. However,  $\kappa$ 's unobservability, requiring pre-known irrelevant donors, makes enforcement challenging. Companion work introduces automatic curation avoiding  $\kappa$ , weights, or factor dimension  $r$  estimates, using alternative geometries.

### Limitations and Extensions

The framework targets single-treated-unit, homoskedastic factor models (Assumptions 2.1-2.5). Extensions to staggered adoption or multi-outcome panels (e.g., VARs) face added issues:

- **Donor Overlap:** Shared controls risk spillover contamination, breaching exclusion.
- **Heterogeneous Structures:** Unit-specific factor loadings may vary relevance per treated unit, necessitating tailored diagnostics.

- **Outcome-Specific Relevance:** Joint assumptions fail if relevance holds for some outcomes but not others, contaminating aggregates.

Formalizing these demands asymptotic expansions for dependence and heterogeneity, deferred to future research.

### Closing Perspective

Panel methods are pivotal in economics, political science, and policy, yet rely on underexamined primitives. We show pre-treatment fit diagnostics falsely reassure via spurious fits when primitives fail.

Results mandate principled protocols over RMSE reliance. By delimiting identifiable elements from finite data, we offer theoretical and practical tools for robust inference. Limits affirm: at  $\kappa = 1$ , residuals detect no violations (Proposition 3.1c); for  $0 < \kappa < 1$ , boundaries diverge (Theorem 4.2). Ex-ante curation minimizes  $\kappa$  theoretically necessarily in dense environments, with screening returns outpacing sample gains (Corollary 4.1). Algorithmic selection (LASSO) on contaminated pools cannot replace structural screening.

Projection residuals' constraints suggest alternative principles. The companion manuscript develops  $\kappa$ -, weight- and dimension-free diagnostics. This work creates the foundation: when methods succeed/fail, and why donor quality trumps size.

### References

- Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490), 493–505.
- Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3), 1203–1227.
- Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11, 727–753.
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12), 4088–4118.
- Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536), 1716–1730.
- Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4), 1229–1279.
- Baik, J., Ben Arous, G., & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5), 1643–1697.
- Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, 116(536), 1789–1803.
- Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4), 1705–1732.
- Botosaru, I., & Ferman, B. (2019). On the role of covariates in the synthetic control method. *The Econometrics Journal*, 22(2), 117–130.

- Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2), 200–230.
- Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772–793.
- Chernozhukov, V., Wüthrich, K., & Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536), 1849–1864.
- Elliott, G., Rothenberg, T. J., & Stock, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica*, 64(4), 813–836.
- El Karoui, N. (2010). The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1), 1–50.
- Ferman, B., & Pinto, C. (2021). Synthetic controls with imperfect matching. *Quantitative Economics*, 12(4), 1197–1246.
- Fernández-Val, I., Freeman, H., & Weidner, M. (2021). Low-rank approximations of nonseparable panel models. *The Econometrics Journal*, 24(2), C40–C77.
- Janson, L., Barber, R. F., & Candès, E. (2017). EigenPrism: Inference for high-dimensional signal-to-noise ratios. *Statistical Science*, 32(4), 606–624.
- Marčenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4), 457–483.
- Newey, W. K., & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5), 1565–1578.
- Pelger, M. (2019). Large-dimensional factor modeling based on high-frequency observations. *Journal of Econometrics*, 208(1), 23–42.
- Roy, O., & Vetterli, M. (2007). The effective rank: A measure of effective dimensionality. *15th European Signal Processing Conference*, 606–610.
- Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.
- Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models* (pp. 80–108). Cambridge University Press.
- Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2), 175–199.
- Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics.
- van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.
- Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1), 57–76.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5), 2183–2202.

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

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

## Appendix A.1 (symbols used)

Symbol	Description
$\lambda_0$	Treated unit's latent loading vector
$\lambda_i$	Unit-specific loading vector for donor i
$f_t$	Common factor vector at time t
$\varepsilon_{it}$	Idiosyncratic noise for unit i at time t
$T_0$	Pre-treatment period length
$T$	Total time periods
$N$	Number of donor units
$r$	Number of latent factors
$Y_{it}(0)$	Untreated potential outcome for unit i at time t
$Y_{it}(1)$	Treated potential outcome for unit i at time t
$y_0$	Pre-treatment outcome vector for treated unit
$X$	Pre-treatment donor matrix
$F$	Factor matrix $[f_1, \dots, f_{T_0}]^T$
$\Lambda$	Donor loadings matrix $[\lambda_1, \dots, \lambda_N]^T$
$E$	Donor error matrix
$\varepsilon_0$	Treated unit's pre-treatment error vector
$G_\Lambda$	Gram matrix $(1/N) \Lambda' \Lambda$
$G_F$	Gram matrix $(1/T_0) F' F$
$\lambda_{\min}$	Minimum eigenvalue
$c_\Lambda$	Lower bound for $\lambda_{\min}(G_\Lambda) > 0$
$c_F$	Lower bound for $\lambda_{\min}(G_F) > 0$
$a_i$	Donor i's alignment with treated unit $[0,1]$
$c$	Lower bound for relevant donors' alignment $> 0$

Symbol	Description
$N_{\text{rel}}$	Number of relevant donors
$N_{\text{irrel}}$	Number of irrelevant donors
$\kappa$	Contamination ratio $\text{rank}(X_{\text{irrel}})/T_0$ or the asymptotic structural saturation parameter
$s$	Effective rank of irrelevant block
$\delta$	Population relevance margin
$\delta^2$	Squared population relevance margin
$\sigma$	Noise standard deviation ( $\sigma^2$ variance)
$\alpha$	Effective saturation (often equiv $\kappa$ )
$H_0$	Null hypothesis $\delta = 0$
$H_1$	Alternative $\delta > 0$
$\Delta_T$	Detection boundary $\sigma (1-\kappa)^{-1/4} \sqrt{(r / T_0)}$
$r$	Projection residual $y_0 - P_X y_0$
$w$	Weight vector
$G_T$	Min of $G_\Lambda$ and $G_F$
$\gamma$	Decay rate $\lambda_{\min} \sim T_0^{-\gamma}$
$v$	Min eigenvalue eigenvector
$Z$	Normal random variable $N(0, \sigma^2)$
$\xi$	Scaled chi-squared $\chi^2_1 / c$
$P_F$	Projector onto $\text{span}(F)$
$Q$	$I - P_F$ (orthogonal complement)
$P_X$	Projector onto $\text{span}(X)$
$\mathcal{S}$	Signal subspace $\text{span}(F)$
$\mathcal{N}$	Noise subspace $\mathcal{S}^\perp$
$\mathcal{U}$	Random irrelevant subspace in $\mathcal{N}$

Symbol	Description
$k_{\text{eff}}$	Effective rank $\text{tr}(P_X)$
$\Sigma_F$	Factor covariance $E[f_t f_t']$
$\mu_T$	Local violation vector
$\ell_T$	Log-likelihood ratio
$R$	Normalized residual $T_0^{-1}$
$\sigma_R$	SD of $R$
$T$	Test statistic $(R - E[R]) / \sigma_R$
$\sigma_X^2$	Donor-based noise variance estimator
$v_N$	Min eigenvalue eigenvector
$z_t$	$X_t' v$ (row projection)
$S_p$	$F P_{\lambda_0}$ (in-span component)
$F_{\delta}$	Structural violation component
$m$	Normalized margin
$\alpha$	Weight on irrelevant in SC
$v$	Normalized irrelevant weights
$C$	Max hull projection in $u_{\delta}$ direction
$u_{\delta}$	Unit direction $F_{\delta} /$
$\alpha^*$	Optimal alpha minimizing SC residual
$K_{\text{eff}}$	Effective contamination in SC
$\mathbb{E}$	Expectation operator
$\perp$	Orthogonal
$\text{tr}$	Trace operator
$\dagger$	Moore-Penrose pseudoinverse
$\oplus$	Direct sum

Symbol	Description
$\Theta_p$	Order in probability
$\approx$	Approximately
$\sim$	Distributed as
$\succ$	Positive definite
$\succeq$	Positive semidefinite
$\circ$	Hadamard product
$\eta_i$	Alignment measure (similar to $a_i$ )
$v$	Asymptotic ratio $N_{\text{rel}}/T_0$
$\hat{f}$	Estimated factors
$\hat{\Lambda}$	Estimated loadings

## Appendix A.3 (proofs of section 3)

### Proposition 3.1 (Unified Asymptotic Decomposition Unstructured and Structured Irrelevance).

Let  $r = y_0 - P_X y_0$  be the projection residual, where  $P_X y_0 = X \hat{w}$  and  $\hat{w} = (X'X)^\dagger X' y_0$  minimize  $\|y_0 - Xw\|_2^2$ . Here  $P_X = X(X'X)^\dagger X'$  denotes the orthogonal projection onto  $\text{span}(X)$ , with  $\dagger$  the Moore–Penrose inverse.

Under Assumptions 2.1–2.5 and Definition 3.2, decompose the donor pool as  $N = N_{\text{rel}} + N_{\text{irrel}}$ . As  $T_0 \rightarrow \infty$  with  $\text{rank}(X_{\text{irrel}})/T_0 \rightarrow \kappa \geq 0$  holding  $N_{\text{rel}}$  fixed or growing sublinearly ( $N_{\text{rel}}/T_0 \rightarrow 0$ ), the normalized squared residual satisfies:

$$\frac{1}{T_0} \|r\|_{2 \rightarrow p}^2 \approx \underbrace{\mathcal{M}(\delta, \mathcal{U}) \cdot \delta^2}_{\text{Attenuated Relevance Violation}} + \underbrace{(1 - \kappa) \cdot \sigma_\epsilon^2}_{\text{Attenuated Noise}}$$

Where  $\delta^2$  is the squared structural relevance violation (Definition 3.1),  $\sigma_\epsilon^2$  is the idiosyncratic variance,  $\mathcal{M}(\delta, \mathcal{U}) \in [0, 1]$  is the signal masking factor (depends on contamination structure), and  $\kappa$  is the contamination ratio. The noise attenuation  $(1 - \kappa)$  is universal across all cases, while the violation attenuation  $\mathcal{M}(\delta, \mathcal{U})$  exhibits case-specific behavior:

**(i) The Precise Formula (Finite Sample Decomposition):** The effective rank decomposes into signal capture and geometric saturation:

$$\frac{k_{\text{eff}}}{T_0} \approx \underbrace{\frac{r}{T_0}}_{\substack{\text{Classical Term } (\rightarrow 0) \\ \text{Signal + Standard} \\ \text{Overfitting}}} + \underbrace{\frac{\min(\text{rank}(X_{\text{irrel}}), T_0 - r)}{T_0}}_{\substack{\text{Structural Term } (\rightarrow \kappa) \\ \text{Spurious Fit}}}$$

(Note: Relevant donors contribute the factor dimension  $r$ . Irrelevant donors fill the remaining noise space  $(T_0 - r)$  up the rank of their outcome matrix  $X_{\text{irrel}}$ .)

It is crucial to distinguish classical overfitting from the spurious fit mechanism: **Standard Overfitting:** Occurs when relevant donors ( $N_{\text{rel}}$ ) fit the idiosyncratic noise of the treated unit. This is a variance issue ( $r/T_0$  term), remediable by increasing  $T_0$  or regularization; **Spurious Fit ( $\kappa$ -Mechanism):** Occurs when irrelevant donors fit the noise or the structural mismatch  $\delta$ . This is a bias issue. The projection loads on dimensions orthogonal to the true factors  $F$ , masking the non-existence of the counterfactual. This form of overfitting persists asymptotically if  $\kappa > 0$  and requires donor purification for dense geometries.

**Noise attenuation** is governed by the rank fraction:  $(1 - \kappa) = 1 - \frac{\text{rank}(X_{\text{irrel}})}{T_0}$ .

**Relevance violation attenuation** depends on contamination structure:

- **Case A (Unstructured/Dense Irrelevance):** When irrelevant donors are mutually uncorrelated with  $\text{rank}(X_{\text{irrel}}) \approx N_{\text{irrel}}$ , the irrelevant subspace  $\mathcal{U}$  is Haar-distributed. By concentration of measure on the Grassmannian:  $\mathcal{M} \rightarrow (1 - \kappa)$  (due to isotropy). Thus, signal and noise are attenuated identically:  $\frac{1}{T_0} \|r\|_{2 \rightarrow p}^2 (\delta^2 + \sigma_\epsilon^2)(1 - \kappa)$ .
- **Case B (Structured/Low-Rank Irrelevance):** When irrelevant donors share orthogonal factors with  $\text{rank}(X_{\text{irrel}}) \ll N_{\text{irrel}}$ , the irrelevant subspace  $\mathcal{U}$  is deterministic. Signal

attenuation depends on the principal angle  $\theta$  between  $\delta$  and  $\mathcal{U}$ :  $\mathcal{M} = \|(I - P_{\mathcal{U}})u_{\delta}\|^2$ , This yields:

- **Best case (for detection):**  $\delta \perp \mathcal{U} \Rightarrow \mathcal{M} = 1$  (no relevance violation masking, full diagnostic power)
- **Worst case (for detection):**  $\delta \in \mathcal{U} \Rightarrow \mathcal{M} = 0$  (complete relevance violation masking)
- **Baseline case (isotropic violation under unstructured irrelevance):** If the violation direction  $u_{\delta}$  is uniformly distributed over the noise space  $N$  and the irrelevant subspace  $U$  is Haar-distributed (as in the dense irrelevance regime), then  $\mathbb{E}[\mathcal{M}] = 1 - \text{rank}(X_{\text{irrel}})/(T_0 - r)$ .

In many economic applications, irrelevant donors may be structured (e.g., agricultural states sharing common regional shocks orthogonal to the treated unit). In this case, the effective saturation  $\kappa$  is determined by the rank of the irrelevant block ( $\text{rank}(X_{\text{irrel}}) \ll N_{\text{irrel}}$ ), preserving more projection-based diagnostic signal.

**(ii) The Asymptotic Limit and Structural Saturation:** As  $T_0 \rightarrow \infty$  with  $N_{\text{rel}}$  fixed, the Classical Term is asymptotically negligible ( $r/T_0 \rightarrow 0$ ), standard overfitting becomes negligible, but the Structural Term persists, defining  $\kappa$ , the contamination ratio or the asymptotic structural saturation parameter:

$$\kappa = \lim_{T_0 \rightarrow \infty} \frac{\min(\text{rank}(X_{\text{irrel}}), T_0 - r)}{T_0}$$

The value of  $\kappa$  depends on the correlation structure of the irrelevant donors:

- **Dense irrelevance:**  $\text{rank}(X_{\text{irrel}}) \approx N_{\text{irrel}} \Rightarrow \kappa = \min(N_{\text{irrel}}/T_0, 1)$
- **Structured irrelevance:**  $\text{rank}(X_{\text{irrel}}) = s \Rightarrow \kappa = \min(s/T_0, 1)$

**(iii) Regime Classification:** The behavior of the residual depends on the saturation level  $\kappa$ :

**(a) Well-Specified Regime ( $\kappa = 0$ ):**

$$k_{\text{eff}} \rightarrow r \Rightarrow \frac{1}{T_0} \|r\|^2 \rightarrow_p (\delta^2 + \sigma_{\epsilon}^2)(1-0)$$

Interpretation: The residual accurately reflects structural bias plus noise. Projection-based pre-treatment fit is a valid diagnostic for relevance violations.

**(b) Partial Spurious Fit ( $0 < \kappa < 1$ ):**

- **Noise:** Always attenuated by factor  $(1-\kappa)$
- **Relevance violation:**
  - Case A Unstructured: Attenuated by  $(1-\kappa)$ , masking is proportional
  - Case B Structured: Attenuated by  $\mathcal{M}(\delta, \mathcal{U}) = \|(I - P_{\mathcal{U}})u_{\delta}\|^2 \in [0,1]$ , masking varies with alignment

**Interpretation:** Irrelevant donors partially saturate the noise subspace. Detection power degrades, but extent depends on whether contamination is fragmented (Case A) or clustered (Case B).

**(c) Spurious Fit Trap ( $\kappa \rightarrow 1$ ):**

$$k_{\text{eff}} \approx T_0 \Rightarrow \frac{1}{T_0} \|r\|^2 \rightarrow_p 0$$

Interpretation: As the effective rank approaches the full outcome dimension, corresponding to maximal-rank contamination in which irrelevant donors span the orthogonal complement, the donor space saturates the outcome space. Both relevance violation and noise are perfectly interpolated, causing pre-treatment RMSE to converge to zero regardless of the magnitude of the relevance violation  $\delta$ . Consequently, fit-based diagnostics become asymptotically uninformative.

**Remark (Convergence and Practical Applicability)**

**(a) Convergence as  $\kappa \rightarrow 1^-$ :** Proposition 3.1 establishes that for any fixed contamination level  $\kappa < 1$ , the residual norm converges to the stated limit as  $T_0 \rightarrow \infty$ . However, we do not claim uniform convergence over all  $\kappa \in [0,1)$  simultaneously. As  $\kappa$  approaches unity, the rate of convergence slows and the approximation quality degrades. This degradation is not a phase transition but a continuous process: diagnostic power erodes smoothly as irrelevant donors increasingly saturate the noise subspace. The limiting behavior as  $\kappa \rightarrow 1^-$  is precisely characterized in Theorem 4.2, which shows that the detection boundary diverges as  $(1-\kappa)^{-1/4}$ , making violations progressively harder to detect even though the residuals remain asymptotically well-defined.

**(b) Finite-Sample Considerations:** The regime classifications in part (iii) rely on asymptotics where the noise subspace dimension  $T_0 - r \rightarrow \infty$ . In empirical applications with short pre-treatment periods, common in comparative case studies (e.g., Abadie et al. 2010 uses  $T_0 = 19$ ), the noise subspace is limited, and high-dimensional random matrix approximations (e.g., Marchenko-Pastur concentration) may exhibit slower convergence. Nonetheless, the finite-sample detection boundary in Theorem 4.2(i-b), which includes an explicit  $\log T_0$  correction, applies exactly for any  $T_0 \geq r$  via sub-Gaussian concentration inequalities. For practical calibration, researchers should interpret regime boundaries as qualitative guides rather than sharp thresholds when  $T_0$  is small.

**(c) Practical Interpretation:** For applied work, the key takeaway is that  $\kappa < 1$  is a necessary condition for diagnostic validity, not a sufficient one. Even when  $\kappa = 0.7$  (seemingly far from saturation), 70% of relevance violations are masked, severely degrading detection power (Theorem 4.2). Conservative practice enforces  $\kappa \leq 0.5$  via ex-ante donor curation, balancing diagnostic sensitivity against sample size. This threshold balances two considerations: (i) preserving at least 50% of diagnostic relevance violation, and (ii) maintaining sufficient sample size for stable weight estimation. Stricter thresholds ( $\kappa \leq 0.3$ ) may be warranted when  $T_0$  is small or violations are expected to be modest.

**Proof.**

**Objective:**

We wish to find the limit of the Normalized Squared Residual  $R$ :

$$R = \frac{1}{T_0} \|r\|_2^2 \text{ as } T_0 \rightarrow \infty$$

**Step 1: Model Setup and Estimator Definition**

First, we define the mathematical objects and the estimator.

**1. Estimator:** The projection estimator  $\hat{y}_0$  minimizes the distance between the treated unit  $y_0$  and the donors  $X$ . The solution is the orthogonal projection.

$$\hat{y}_0 = P_X y_0$$

where  $P_X = X(X^\top X)^\dagger X^\top$  is the unique orthogonal projection matrix onto  $\text{span}(X)$ .

**2. Residual:** The residual is the difference between the observed outcome and the projection.

$$r = y_0 - \hat{y}_0 = (I - P_X)y_0$$

### 3. Data Structure:

- $y_0 = F\lambda_0 + \epsilon_0$  (Assumption 2.1).
- $X = [X_{\text{rel}}, X_{\text{irrel}}]$ .
- Relevant donors span the factors:  $\text{span}(X_{\text{rel}}) \approx \text{span}(F)$ .
- Irrelevant donors are orthogonal to factors:  $\text{span}(X_{\text{irrel}}) \perp \text{span}(F)$ .

### Step 2: Geometric Decomposition of the Treated Unit

We must split the treated unit vector  $y_0$  into a **Signal Component** (spanned by factors) and a **Noise/Violation Component** (orthogonal to factors).

#### Definition (True Factor Projectors):

Let  $F$  be the matrix of true factors.

- $P_F = F(F^\top F)^{-1}F^\top$ : Projector onto the Signal Space  $\mathcal{S}$ .
- $Q = I - P_F$ : Projector onto the Noise Space  $\mathcal{N}$  (Orthogonal Complement).

#### Decomposition:

Apply the identity matrix  $I = P_F + Q$  to  $y_0$ :

$$\begin{aligned} y_0 &= (P_F + Q)(F\lambda_0 + \epsilon_0) \\ &= P_F(F\lambda_0) + Q(F\lambda_0) + P_F(\epsilon_0) + Q(\epsilon_0) \end{aligned}$$

#### Analysis of Terms:

1.  $P_F(F\lambda_0)$ : Since  $F\lambda_0$  is already in the span of  $F$ , projecting it doesn't change it.

$$= F\lambda_0$$

2.  $Q(F\lambda_0)$  (**The Structural Violation**): This is the part of the treated unit's signal that is *orthogonal* to the donor factors.

- *Why  $\sqrt{T_0}$ ?* The Euclidean norm squared of a persistent relevance violation grows linearly with time ( $T_0$ ). To define a stable population parameter  $\delta^2$  (Mean Squared Error), we must normalize the vector.
- We define  $\delta_T$  such that  $\sqrt{T_0}\delta_T = Q(F\lambda_0)$ .
- Assuming the violation component satisfies a LLN so that  $\|Q(F\lambda_0)\|^2/T_0 \rightarrow \delta^2$ .
- Thus,  $\|\sqrt{T_0}\delta_T\|^2/T_0 = \|\delta_T\|^2 \rightarrow \delta^2$ .

3.  $P_F(\epsilon_0)$ : This is noise projected onto a low-dimensional space ( $r \ll T_0$ ). By the Law of Large Numbers, its energy is negligible ( $O_p(r/T_0) \rightarrow 0$ ). We ignore it for the leading order derivation.

4.  $Q(\epsilon_0)$ : This is the noise remaining in the high-dimensional space.

#### Resulting Equation for $y_0$ :

$$y_0 = F\lambda_0 + \sqrt{T_0}\delta_T + Q\epsilon_0$$

$$\underbrace{\quad}_{\in \mathcal{S}} \quad \underbrace{\quad}_{\in \mathcal{N}} \quad \underbrace{\quad}_{\in \mathcal{N}}$$

### Step 3: Geometric Decomposition of the Donor Span

Now we determine what the estimator  $P_X$  actually projects onto. We use **Random Matrix Theory** to approximate the subspaces.

#### 1. Relevant Donors ( $X_{\text{rel}}$ ):

Since they load on the factors, their span approximates the Signal Space  $\mathcal{S}$ .

- **Theorem Used: Davis-Kahan sin  $\Theta$  Theorem (1970).** This bounds the angle between the perturbed subspace (observed donors) and the true subspace (factors). For high Signal-to-Noise Ratio, the angle  $\rightarrow 0$ .
- **Result:**  $\text{span}(X_{\text{rel}}) \approx \text{span}(F) = \mathcal{S}$ , under the eigenvalue separation implied by Assumption 2.1.

#### 2. Irrelevant Donors ( $X_{\text{irrel}}$ ):

These lie in the Noise Space  $\mathcal{N}$ . Let  $\mathcal{U} = \text{span}(QX_{\text{irrel}})$  be the specific subspace they span.

We need the **dimension** (rank) of  $\mathcal{U}$  to know how much noise they capture.

- **Theorem Used: Marchenko-Pastur Law (1967).**

Under standard nondegeneracy and moment conditions ensuring MP behavior for the sample covariance  $\frac{1}{T_0} X_{\text{irrel}}^\top X_{\text{irrel}}$  (e.g., independent rows, or weakly dependent with suitable mixing, with uniformly bounded fourth moments), the irrelevant block is full rank so that whenever  $N_{\text{irrel}}/T_0 \leq \kappa^- < 1$ , the sample covariance is full rank with probability tending to one:

$$\lambda_{\min} \rightarrow (1 - \sqrt{\frac{N_{\text{irrel}}}{T_0}})^2 > 0$$

- **Result:** Assume the population covariance of  $X_{\text{irrel}}$  has eigenvalues bounded away from zero. The dimension of  $\mathcal{U}$  is exactly  $k_{\text{eff}} = \text{rank}(X_{\text{irrel}}) \approx \min(N_{\text{irrel}}, T_0 - r)$ .

#### 3. Total Projector ( $P_X$ ):

We invoke High-Dimensional Geometry.

- **Theorem Used: Vershynin (2018, Thm 4.6.1).** Two high-dimensional random vectors from orthogonal distributions are approximately orthogonal.
- Since  $\mathcal{S} \perp \mathcal{N}$ , standard random subspace incoherence results imply that the principal angles between  $\text{span}(X_{\text{rel}})$  and  $\text{span}(X_{\text{irrel}})$  converge to  $\pi/2$ , so that  $\|P_{X_{\text{rel}}}P_{X_{\text{irrel}}}\| = o_p(1)$ ,
- and cross-projection terms are asymptotically negligible.
- Under Assumption 2.1 and sub-Gaussianity of  $X_{\text{irrel}}$ , standard results on random subspace incoherence imply  $\|P_X - (P_F + P_U)\| = o_p(1)$  in operator norm.
- Consequently,  $P_{X_{\text{rel}}} \rightarrow_p P_F$  and  $P_X \rightarrow_p P_F + P_U$ , with convergence in operator norm (or Frobenius norm).

where  $P_F$  projects onto Signal, and  $P_U$  projects onto the Irrelevant Subspace inside Noise.

### Step 4: The Residual Expansion

We substitute the decomposition of  $P_X$  (Step 3) and  $y_0$  (Step 2) into the residual definition.

$$r = (I - P_X)y_0 \approx (I - (P_F + P_U))y_0$$

### Algebraic Expansion:

$$r \approx (I - P_F - P_U)(\underbrace{F\lambda_0}_{\text{Signal}} + \underbrace{\sqrt{T_0}\delta_T + Q\epsilon_0}_{\text{Violation/Noise}})$$

### Annihilation:

1. **Signal Term:**  $(I - P_F - P_U)F\lambda_0$ .

Since  $F\lambda_0 \in \mathcal{S}$ ,  $P_F(F\lambda_0) = F\lambda_0$ .

$$= F\lambda_0 - F\lambda_0 - 0 = 0$$

(The relevant donors define the signal space and successfully explain the signal).

2. **Violation/Noise Terms:**

These vectors live in  $\mathcal{N}$ , so  $P_F(\dots) = 0$ .

However,  $P_U$  acts on them.

The operator becomes:

$$(I - 0 - P_U) = (I - P_U)$$

### Final Residual Vector:

$$r \approx (I - P_U)(\sqrt{T_0}\delta_T + Q\epsilon_0)$$

*Interpretation:* The residual is composed of the Structural Violation and the Noise, **minus** whatever part the Irrelevant Donors ( $P_U$ ) managed to capture (spuriously).

### Step 5: The Norm Calculation (The Trace Trick)

We compute the normalized squared norm  $R = \frac{1}{T_0} \|r\|_2^2$ .

$$R = \frac{1}{T_0} \|(I - P_U)\sqrt{T_0}\delta_T + (I - P_U)Q\epsilon_0\|^2$$

Expand the square  $\|A + B\|^2 = \|A\|^2 + \|B\|^2 + 2A^\top B$ :

$$R = \underbrace{\frac{1}{T_0} \|(I - P_U)\sqrt{T_0}\delta_T\|^2}_{\text{Relevance Violation Term}} + \underbrace{\frac{1}{T_0} \|(I - P_U)Q\epsilon_0\|^2}_{\text{Noise Term}} + \underbrace{\text{Cross Term}}_{\xrightarrow{\sim} 0}$$

Cross term:  $\frac{2}{T_0} (\sqrt{T_0}\delta_T)^\top (I - P_U)Q\epsilon_0$ . Since  $\epsilon_0$  is independent of  $\delta$  (Assumption 2.2), this has mean zero and variance  $\text{Var}(\delta_T^\top \epsilon_0 / \sqrt{T_0}) = \|\delta_T\|^2$  and  $\sigma_\epsilon^2/T_0 = O(1/T_0) \rightarrow 0$ .

(The cross term is  $o_p(1)$  by Cauchy-Schwarz and concentration of sub-Gaussian quadratic forms, since  $\|(I - P_U)\delta_T\| = O(1)$  and  $\|(I - P_U)Q\epsilon_0\| = O_p(\sqrt{T_0})$ .)

### 5.1 Analyzing the Noise Term (Universal Attenuation)

We calculate the expected value of the Noise Term:

$$\mathbb{E}[N_{term}] = \frac{1}{T_0} \mathbb{E}[\epsilon_0^\top Q(I - P_U)^\top (I - P_U) Q \epsilon_0]$$

**Algebraic Simplification:**

1.  $P_U$  is a subspace of  $Q$ . Thus  $Q(I - P_U) = Q - P_U$ .
2. Let  $M = Q - P_U$ . This is a projector, so  $M^\top M = M$ .
3. The term becomes  $\epsilon_0^\top M \epsilon_0$ .

**The Trace Trick:**

For  $\epsilon_0 \sim (0, \sigma^2 I)$ ,  $\mathbb{E}[\epsilon^\top M \epsilon] = \sigma^2 \text{tr}(M)$ .

$$\text{tr}(M) = \text{tr}(Q) - \text{tr}(P_U)$$

**Using Dimensions (From Marchenko-Pastur in Step 3):**

- $\text{tr}(Q) = \dim(\mathcal{N}) = T_0 - r$ .
- $\text{tr}(P_U) = \dim(\mathcal{U}) = k_{\text{eff}}$ .

**Result:**

$$\mathbb{E}[N_{term}] = \frac{\sigma_\epsilon^2}{T_0} (T_0 - r - k_{\text{eff}}) = \sigma_\epsilon^2 \left( 1 - \frac{r}{T_0} - \frac{k_{\text{eff}}}{T_0} \right)$$

As  $T_0 \rightarrow \infty$ , this converges to  $(1 - \kappa) \sigma_\epsilon^2$ .

*Theorem Used: Hanson-Wright Inequality (Rudelson & Vershynin, 2013)* ensures the random quadratic form concentrates tightly around this expected value.

**5.2 Analyzing the Relevance Violation Term (Case Specific)**

$$V_{term} = \| (I - P_U) \delta_T \|^2$$

Let  $u_{\delta_T} = \delta_T / \|\delta_T\|$  be the unit vector of the violation.

$$= \|\delta_T\|^2 \cdot \| (I - P_U) u_{\delta_T} \|^2$$

**Case A: Dense Irrelevance (Random Subspace)**

- $X_{\text{irrel}}$  is random noise.
- **Theorem Used: Rotational Invariance of Gaussian Matrices.** The subspace  $\mathcal{U}$  is uniformly distributed (Haar Measure) on the Grassmannian manifold.
- Projecting a fixed vector onto a random subspace of size  $\kappa$  captures exactly fraction  $\kappa$  of the energy in expectation.
- $\mathbb{E}[\|P_U u_{\delta_T}\|^2] = \frac{\dim(\mathcal{U})}{\dim(\mathcal{N})} = \kappa$
- Therefore, the remaining energy is  $(1 - \kappa)$ .
- **Result:**  $(1 - \kappa) \delta^2$ .

**Case B: Structured Irrelevance (Fixed Subspace)**

- $X_{\text{irrel}}$  shares fixed factors.  $\mathcal{U}$  is a deterministic subspace.
- We cannot use the Haar measure average.
- We define the geometric masking factor  $\mathcal{M}(\delta_T, \mathcal{U}) = \| (I - P_U) u_{\delta_T} \|^2$ . Equivalently, if  $\theta$  is the principal angle between  $\delta_T$  and  $\mathcal{U}$ , then  $\mathcal{M} = \sin^2(\theta)$  (Pythagorean theorem in subspace decomposition).

- *Result:*  $\mathcal{M}(\delta, \mathcal{U})\delta^2$ .

### Step 6: Final Result

Combining Step 5.1 (Noise) and Step 5.2 (Relevance Violation):

$$\frac{1}{T_0} \|r\|^2 \rightarrow_p \underbrace{\mathcal{M}(\delta, \mathcal{U}) \cdot \delta^2}_{\text{Attenuated Relevance Violation}} + (1 - \kappa) \cdot \underbrace{\sigma_\epsilon^2}_{\text{Attenuated Noise}}$$

This completes the derivation.  $\square$

### Remark 3.3 (Regularization and the Spurious Fit Trap)

While Proposition 3.1 characterizes spurious fit under ordinary least squares, practitioners often employ regularized estimators (Ridge, LASSO, Elastic Net) to improve stability or prediction accuracy. We now clarify how regularization modifies the effective dimension  $k_{\text{eff}}$  but does not fundamentally alter the spurious fit mechanism or restore diagnostic validity when  $\kappa > 0$ .

#### Ridge Regression: Modified Attenuation Without Identification Recovery

Consider the ridge estimator with penalty parameter  $\lambda > 0$ :

$$\hat{w}^\lambda = (X'X + \lambda I)^{-1} X'y_0$$

The ridge-smoothed prediction is  $\hat{y}_0^\lambda = S_\lambda y_0$  where  $S_\lambda = X(X'X + \lambda I)^{-1} X'$  is the ridge smoother matrix. The effective dimension becomes:

$$k_\lambda = \text{tr}(S_\lambda) = \sum_{j=1}^{\min(N, T_0)} \frac{\xi_j}{\xi_j + \lambda}$$

where  $\xi_j$  are the eigenvalues of  $X'X/T_0$ .

Under proportional asymptotics with  $\text{rank}(X_{\text{irrel}})/T_0 \rightarrow \kappa$  and  $\lambda/T_0 \rightarrow \bar{\lambda} \in [0, \infty)$ , the Marchenko-Pastur law (Marčenko & Pastur 1967; El Karoui 2010) implies:

$$\frac{k_\lambda}{T_0} \rightarrow c(\bar{\lambda}, \kappa) = \int_0^\infty \frac{t}{t + \bar{\lambda}} dF_\kappa(t)$$

where  $F_\kappa$  is the limiting spectral distribution of  $X'X/T_0$  under aspect ratio  $\kappa$ .

#### Key properties:

1.  $c(0, \kappa) = \min(\kappa, 1)$  recovers the OLS limit from Proposition 3.1
2.  $c(\bar{\lambda}, \kappa)$  is strictly decreasing in  $\bar{\lambda}$  for fixed  $\kappa$ , with  $c(\infty, \kappa) = 0$
3. For any  $\bar{\lambda} > 0$ :  $c(\bar{\lambda}, \kappa) < c(0, \kappa) = \min(\kappa, 1)$

The ridge residual satisfies:

$$\frac{1}{T_0} \|y_0 - \hat{y}_0^\lambda\|^2 \rightarrow (\delta^2 + \sigma_\epsilon^2)(1 - c(\bar{\lambda}, \kappa)) + o_p(1)$$

### Regime (a): Well-Specified ( $\kappa = 0$ )

When all donors are relevant ( $N_{\text{irrel}} = 0$ ), ridge provides:

- **OLS:**  $\frac{1}{T_0} \|r\|^2 \rightarrow \delta^2 + \sigma_\epsilon^2$
- **Ridge:**  $\frac{1}{T_0} \|r^\lambda\|^2 \rightarrow (\delta^2 + \sigma_\epsilon^2)(1 - c(\bar{\lambda}, 0))$

Since  $c(\bar{\lambda}, 0) > 0$  when  $\bar{\lambda} > 0$ , the factor  $(1 - c(\bar{\lambda}, 0))$  inflates residuals even when relevance holds perfectly ( $\delta = 0$ ). This penalty-induced bias degrades diagnostic power by increasing false positive rates.

**Implication:** In well-specified settings, regularization is counterproductive for diagnostics. Use OLS residuals.

### Regime (b): Partial Spurious Fit ( $\kappa \in (0, 1)$ )

Ridge modifies the attenuation factor but does not eliminate the fundamental masking problem:

- **OLS:**  $\frac{1}{T_0} \|r\|^2 \rightarrow (\delta^2 + \sigma_\epsilon^2)(1 - \kappa)$
- **Ridge:**  $\frac{1}{T_0} \|r^\lambda\|^2 \rightarrow (\delta^2 + \sigma_\epsilon^2)(1 - c(\bar{\lambda}, \kappa))$

Since  $c(\bar{\lambda}, \kappa) < \kappa$  for  $\bar{\lambda} > 0$ , ridge **increases** the residual relative to OLS:  $(1 - c(\bar{\lambda}, \kappa)) > (1 - \kappa)$ . This superficially appears beneficial—less masking of  $\delta^2$ .

**However, this improvement is fundamentally limited:**

1. **Penalty-bias confounding:** The residual conflates structural mismatch ( $\delta^2$ ) and penalty-induced bias. Both signal and noise are scaled by the unknown factor  $(1 - c(\bar{\lambda}, \kappa))$ . Without independent knowledge of  $c(\bar{\lambda}, \kappa)$ , this cannot be decomposed into identifiable components.
2. **Circularity in calibration:** Computing  $c(\bar{\lambda}, \kappa)$  requires knowing  $\kappa = \text{rank}(X_{\text{irrel}})/T_0$ , which requires identifying which donors are irrelevant—precisely what diagnostics aim to determine.
3. **Data-driven tuning negates the improvement:** If  $\lambda$  is chosen via cross-validation to minimize pre-treatment prediction error, then as  $T_0 \rightarrow \infty$ :  $\lambda_{\text{CV}}/T_0 \rightarrow 0$  implies  $c(\bar{\lambda}_{\text{CV}}, \kappa) \rightarrow \kappa$ . This occurs because when  $\kappa \in (0, 1)$ , irrelevant donors provide genuine predictive value for interpolating noise in the  $(T_0 - r)$ -dimensional orthogonal subspace. Penalizing their weights reduces in-sample fit without improving structural alignment, so cross-validation selects smaller penalties. The residual collapses back toward the OLS limit.
4. **Fixed penalty: diagnostic benefit with estimation cost:** If  $\lambda$  is held fixed with  $\bar{\lambda} > 0$ :
  - **Diagnostic benefit:** Residual  $(1 - c(\bar{\lambda}, \kappa)) > (1 - \kappa)$  preserves more signal
  - **Estimation cost:** Introduces bias  $\approx \bar{\lambda} \|E[w_0]\|^2$

- **Inference validity:** Standard errors remain invalid unless adjusted for both penalty bias and attenuation factor

**Quantitative assessment:** Suppose  $\kappa = 0.6$  and  $\bar{\lambda} = 0.1$ . From Marchenko-Pastur theory,  $c(0.1, 0.6) \approx 0.45$ . Then:

- OLS residual:  $(1 - 0.6)(\delta^2 + \sigma_\epsilon^2) = 0.4(\delta^2 + \sigma_\epsilon^2)$ , masking 60% of signal
- Ridge residual:  $(1 - 0.45)(\delta^2 + \sigma_\epsilon^2) = 0.55(\delta^2 + \sigma_\epsilon^2)$ , masking 45% of signal

Ridge improves signal preservation by factor 1.375 (37.5% improvement). However, the detection boundary improves only from  $1.58\sigma_\epsilon T_0^{-1/4}$  to  $1.35\sigma_\epsilon T_0^{-1/4}$ —a 15% reduction. As  $\kappa \rightarrow 1^-$ , the detection boundary still diverges even with fixed  $\bar{\lambda} > 0$ .

### Regime (c): Spurious Fit Trap ( $\kappa > 1$ )

When irrelevant donors exceed pre-treatment periods:

- **OLS:**  $\frac{k_{OLS}}{T_0} \rightarrow 1$ , so  $\frac{1}{T_0} \|r\|^2 \rightarrow 0$  regardless of  $\delta$
- **Ridge:**  $\frac{k_\lambda}{T_0} \rightarrow c(\bar{\lambda}, \kappa) \in (0, 1)$  for  $\bar{\lambda} > 0$

Ridge prevents complete numerical collapse, yielding:

$$\frac{1}{T_0} \|r^\lambda\|^2 \rightarrow (\delta^2 + \sigma_\epsilon^2)(1 - c(\bar{\lambda}, \kappa))$$

This superficially appears to "solve" the spurious fit trap. **This is illusory:**

- The residual reflects **penalty-induced bias**, not structural alignment
- Both  $\delta^2$  and  $\sigma_\epsilon^2$  are scaled by the same unknown factor
- With CV-tuned  $\lambda$ :  $\lambda_{CV}/T_0 \rightarrow 0$ , driving  $c(\bar{\lambda}_{CV}, \kappa) \rightarrow 1$  and collapsing residuals to zero as in OLS

**Implication:** When  $\kappa > 1$ , ridge with data-driven tuning replicates the OLS failure. Ridge with fixed  $\bar{\lambda}$  produces non-zero but penalty-determined residuals uninformative about  $\delta$ .

### LASSO and Elastic Net: When Sparsity Can and Cannot Help

For LASSO ( $\ell^1$  penalty) and Elastic Net (combined  $\ell^1 + \ell^2$ ), the behavior depends critically on whether the true donor weights admit a sparse representation.

#### Case 1: True Sparsity (Uncommon in Factor Models)

Suppose the treated unit's counterfactual can be represented using only  $s \ll N$  donors:

$$\lambda_0 = \sum_{i \in S_0} w_{0i} \lambda_i, |S_0| = s \ll N$$

where  $S_0$  is the true support. This occurs when:

- The treated unit shares factors with only a small subset of donors
- Many donors operate in fundamentally different economic regimes

- Institutional features create natural clustering with sparse between-cluster relevance

**When LASSO succeeds:** If signal strength satisfies  $\min_{i \in S_0} |w_{0i}| \gg \sigma_\epsilon \sqrt{\log N/T_0}$  and the penalty is properly tuned to target sparsity (not prediction error), LASSO can successfully recover the sparse support  $\hat{S} \approx S_0$  with high probability (Bühlmann & van de Geer 2011).

In this case, LASSO effectively **purifies the donor pool** by excluding irrelevant units. If the selected model satisfies:

$$(s + s_{\text{spurious}})/T_0 < 1$$

where  $s_{\text{spurious}}$  is the number of falsely included irrelevant donors, then the effective  $\kappa_{\text{eff}} = (s + s_{\text{spurious}})/T_0 < 1$  and diagnostics can succeed in the partial spurious fit regime.

#### Necessary conditions for LASSO success:

1. **Oracle knowledge:** Sparsity must actually hold in the population (untestable without external information)
2. **Sufficient signal:** Minimum non-zero weight exceeds  $\sigma_\epsilon \sqrt{\log N/T_0}$ , the LASSO detection threshold
3. **Proper tuning:** Penalty  $\lambda_1$  must target sparsity (maximize true negatives among irrelevant donors) rather than minimize in-sample prediction error

**Practical limitation:** However, this reliance on sparsity creates an epistemological trap: researchers cannot verify whether the true model is sparse based solely on pre-treatment fit. A dense factor model with high contamination ( $\kappa \approx 1$ ) can generate a "sparse" solution path indistinguishable from a true sparse model, as the algorithm greedily selects irrelevant donors that happen to align spuriously with the noise realization  $\epsilon_0$ .

Furthermore, cross-validation introduces a circularity analogous to ridge regression. When  $N/T_0 > 1$ , selecting  $\lambda_1$  to minimize RMSE typically drives the penalty toward zero and the active set size  $|\hat{S}|$  toward  $N$ , effectively negating donor purification. Tuning specifically for sparsity, rather than prediction, requires auxiliary information about which donors are relevant, which is precisely what the diagnostic aims to determine. Consequently, relying on LASSO to validate identification is circular: it succeeds only if the researcher assumes a priori that the data generating process favors sparsity.

#### Case 2: Dense Representations (Typical in Factor Models)

When many donors have non-zero weights in the population—the typical case under low-rank factor structures where  $\lambda_0 = \sum_{i=1}^N w_{0i} \lambda_i$  with most  $w_{0i} \neq 0$ —LASSO and Elastic Net cannot achieve effective donor purification.

**Elastic Net asymptotics under density:** Consider Elastic Net with penalties  $\lambda_1, \lambda_2 > 0$ :

$$\hat{w}^{\text{EN}} = \arg \min_w \frac{1}{2T_0} \|y_0 - Xw\|^2 + \lambda_1 \|w\|_1 + \frac{\lambda_2}{2} \|w\|_2^2$$

Under the conditions:

1. The true loading  $\lambda_0$  admits no sparse representation:  $\min_{|S| \leq s} \inf_{w: \text{supp}(w) \subseteq S} \|\lambda_0 - \Lambda' w\| > 0$  for all  $s \ll N$
2. Penalty parameters satisfy  $\lambda_1 = o(T_0^{-1/2})$  and  $\lambda_2/T_0 \rightarrow \bar{\lambda} \in (0, \infty)$

The Elastic Net active set diverges:  $|S_{\text{EN}}|/T_0 \rightarrow \kappa_{\text{active}} > 0$ , where  $\kappa_{\text{active}}$  depends on the interplay between  $\lambda_1, \lambda_2$ , and the donor correlation structure.

In this regime, the effective dimension satisfies:

$$\frac{k_{\text{EN}}}{T_0} = \frac{\text{tr}(S_{\lambda_1, \lambda_2})}{T_0} \rightarrow c(\bar{\lambda}, \kappa_{\text{active}}) + o(1)$$

where  $S_{\lambda_1, \lambda_2}$  is the Elastic Net smoother matrix restricted to the active set. The residual behavior becomes:

$$\frac{1}{T_0} \|y_0 - \hat{y}_0^{\text{EN}}\|^2 \rightarrow (\delta^2 + \sigma_\epsilon^2)(1 - c(\bar{\lambda}, \kappa_{\text{active}}))$$

This is **analogous to ridge** with aspect ratio  $\kappa_{\text{active}}$  determined by the selected donors.

**Key insight:** The  $\ell^1$  component performs variable selection among donors but:

- Cannot distinguish relevant from irrelevant donors based solely on pre-treatment fit when both span the noise space
- Selects donors based on correlation with residuals, not true factor alignment
- When  $N/T_0 > 1$ , typically selects  $|S_{\text{EN}}| \approx \min(N, cT_0)$  for some  $c \in (0.7, 1)$ , yielding  $\kappa_{\text{active}} \approx c \cdot N/T_0$

If  $N/T_0 > 1.43$  and  $c = 0.7$ , then  $\kappa_{\text{active}} \approx 1$ , placing Elastic Net back in the spurious fit trap.

**Implication:** In dense factor models—the setting where  $\lambda_0$  is a non-trivial linear combination of many donor loadings—Elastic Net inherits the spurious fit problem from ridge, with the active set size determining the effective  $\kappa_{\text{active}}$ . Unless genuine sparsity holds and penalties are tuned to exploit it (both untestable), Elastic Net does not overcome the geometric saturation mechanism.

### When Does Sparsity-Inducing Regularization Help?

LASSO and Elastic Net can improve diagnostics **only when**:

1. **True sparsity holds:**  $|\{i: w_{0i} \neq 0\}| \ll N$  in population
2. **Signal is strong:** Minimum non-zero weight  $\gg \sigma_\epsilon \sqrt{\log N/T_0}$
3. **External guidance exists:** Auxiliary information (covariates, domain knowledge) informs penalty tuning toward sparsity rather than prediction accuracy
4. **Resulting pool satisfies:**  $(s + s_{\text{spurious}})/T_0 < 1$  after selection

**In practice, these conditions rarely hold simultaneously for panel data applications because:**

- Factor models induce dense representations (all units load on common factors)
- Weak signal strengths (moderate  $w_{0i}$ ) make detection difficult
- Cross-validation optimizes prediction, not donor relevance
- Even aggressive penalization when  $N/T_0 > 1$  often yields  $|S_{EN}|/T_0 \geq 1$

### Unified Practical Guidance Across Regimes

Regularization addresses **classical overfitting** (fitting idiosyncratic noise from relevant donors when  $N_{\text{rel}} \gg T_0$ ) by trading bias for variance. It does **not** address **spurious fit from irrelevant donors** because the geometric mechanism, irrelevant units spanning the orthogonal subspace, operates independently of weight magnitudes or selection patterns.

#### Regime-specific recommendations:

##### 1. $\kappa = 0$ (well-specified):

- Regularization unnecessary for diagnostics; use OLS residuals
- Ridge/LASSO introduce penalty bias that degrades diagnostic power

##### 2. $\kappa \in (0,1)$ (partial spurious fit):

- Ridge with **fixed**  $\bar{\lambda} \in (0.1, 0.5)$  provides modest diagnostic power improvement (15-40% reduction in detection threshold) at cost of penalty bias, for detecting violations but not estimation validity (removing bias).
- **Report both OLS and ridge residuals:** If  $\|r^\lambda\|^2 / \|r^{\text{OLS}}\|^2 > 1.5$ , suspect substantial irrelevant contamination
- **Do not use CV-tuned penalties:** Data-driven tuning drives  $\bar{\lambda} \rightarrow 0$  or  $\lambda_1 \rightarrow 0$ , negating diagnostic benefit
- **LASSO may help if sparsity holds:** But requires external guidance for tuning; cross-validation on pre-treatment data selects dense models
- **Prioritize donor purification:** Reducing  $N_{\text{irrel}}$  provides unambiguous improvement

##### 3. $\kappa > 1$ (spurious fit trap):

- Ridge prevents numerical explosion but produces penalty-determined residuals
- With CV tuning, residuals collapse to zero as in OLS
- With fixed  $\bar{\lambda}$ , residuals reflect unknown attenuation factor
- LASSO typically selects  $|S| \geq T_0$  when  $N \gg T_0$ , failing to purify
- **Diagnostics fail regardless of regularization strategy**

**Fundamental takeaway:** The spurious fit trap is a **geometric phenomenon**, not a statistical one. Regularization modulates the attenuation factor but does not eliminate it. The only reliable remedy is **ex-ante donor curation** ensuring  $\text{rank}(X_{\text{irrel}})/T_0 \rightarrow 0$ , not ex-post regularization after including all available donors.

#### Recommended practice:

- Screen donors using auxiliary information (institutional comparability, economic similarity, covariate balance) before estimation
- Enforce  $N/T_0 < 1$  as necessary condition, targeting  $N/T_0 \leq 0.5$  to preserve diagnostic power
- Use regularization to address weak conditioning (Regime II from Section 5), not to validate identification in presence of spurious fit
- If using LASSO/Elastic Net, validate that selected donor sets satisfy  $(|S| + \text{expected false positives})/T_0 < 1$ ; otherwise, diagnostics remain unreliable

### Extension: Generalized Proposition 3.1 (Structured Irrelevance)

Suppose the irrelevant donors follow a structured factor model in the noise subspace:  $X_j = F_\perp \lambda_j + \epsilon_j$  for irrelevant  $j$ , where  $F_\perp$  is  $T_0 \times s$  with  $\text{rank}(F_\perp) = s$ ,  $F^\top F_\perp = 0$ ,  $\|\lambda_j\|/\sqrt{s} \geq c > 0$ , and  $\epsilon_j$  sub-Gaussian. Then

$$\frac{\|r\|^2}{T_0} = (\delta^2 + \sigma_\epsilon^2)(1 - \kappa_{\text{eff}}) + o_p(1),$$

where  $\kappa_{\text{eff}} = \dim(\text{span}(P_\perp X_i))/(T_0 - r)$  is the effective contamination ratio and  $X_i$  collects the irrelevant-donor columns. Regimes are defined with respect to  $\kappa_{\text{eff}}$  (otherwise identical to Proposition 3.1).

When the irrelevant donors are unstructured ( $s = 0$  or negligible),  $\kappa_{\text{eff}} \approx \kappa$  and the original result is recovered.

### Full Proof of Generalized Proposition 3.1

**Notation and Setup.** Treated unit indexed by 0, donors by  $j = 1, \dots, N$ , pre-treatment periods  $T_0$ . Untreated outcomes satisfy  $Y_{jt}(0) = \lambda_j^\top f_t + \epsilon_{jt}$ . Pre-treatment vectors:  $Y_0 = F\lambda_0 + e_0$ ,  $X = [X_1 \dots X_N]$  with  $X_j = F\lambda_j + e_j$ . Partition: relevant donors ( $N_r$ , high SNR  $\|\lambda_j\|/\sqrt{r} \geq c > 0$ ) and irrelevant donors ( $N_i$ , low SNR w.r.t.  $F$ ). Relevance margin  $\delta^2 = \min_w \|E[Y_0] - Xw\|^2/T_0$ . Residual  $r = (I - P_X)Y_0$ .

**Structured Irrelevance Assumption.** For irrelevant donors:  $X_j = F_\perp \lambda_j + \epsilon_j$ ,  $F^\top F_\perp = 0$ ,  $\text{rank}(F_\perp) = s$  fixed (or growing slowly), full-rank loadings matrix  $\Lambda_i (s \times N_i)$ ,  $\epsilon_j$  sub-Gaussian with variance proxy  $\sigma_\epsilon^2$  (Assumption 2.3).

**Geometric Decomposition.** Outcome space decomposes as signal subspace  $S = \text{col}(F)$  ( $\dim r$ ) and orthogonal complement  $\perp S$  ( $\dim T_0 - r$ ). Projectors  $P_F$  and  $P_\perp = I - P_F$ .

- Relevant donors:  $P_F X_j \approx X_j$  w.h.p. (Vershynin 2018, Thm 4.6.1), so  $\text{span}(\{X_j\}_{\text{relevant}}) \approx S$  and  $P_\perp X_j = o_p(\sqrt{T_0})$ .
- Irrelevant donors:  $P_F X_j = o_p(\sqrt{T_0})$  (orthogonality),  $P_\perp X_j \approx F_\perp \lambda_j + P_\perp \epsilon_j$ . Thus  $X_i = F_\perp \Lambda_i + E_i$ .

The span in the noise subspace is  $\text{col}(P_\perp X_i) \approx \text{col}(F_\perp)$ : the error term  $P_\perp E_i$  perturbs singular values but does not increase rank beyond sw.h.p. (Davis–Kahan sin  $\Theta$  theorem applied to the perturbation  $E_i$ ; bounded moments suffice). Hence  $\dim(\text{col}(P_\perp X_i)) = \min(s, T_0 - r)$  w.h.p. (assuming  $\text{rank}(\Lambda_i) = s$ ).

Total donor span:  $\text{col}(X) \approx S \oplus U$  where  $U \subseteq \perp S$  and  $\dim(U) = s' = \min(s, T_0 - r)$ .

### Residual Expansion.

$$r = (I - P_X)Y_0, \|r\|^2/T_0 = \|(I - P_X)Y_0\|^2/T_0.$$

Decompose  $Y_0 = P_F Y_0 + P_\perp Y_0$ . Because  $P_X \approx P_F + P_U$  (relevant donors cover  $S$ , irrelevant donors cover  $U$ ),

$$(I - P_X)Y_0 \approx (I - P_F - P_U)P_\perp Y_0 = (I - P_U)P_\perp Y_0$$

(up to  $o_p(\sqrt{T_0})$  cross terms that vanish after normalization). Therefore

$$\frac{\|r\|^2}{T_0} = \frac{\|(I - P_U)P_\perp Y_0\|^2}{T_0} + o_p(1).$$

The operator  $I - P_U$  annihilates exactly the fraction  $s'/(T_0 - r) = \kappa_{\text{eff}}$  of the noise subspace. Under the factor model,

$$\mathbb{E}[\|P_\perp Y_0\|^2/T_0] = \delta^2 + \sigma_\epsilon^2$$

(structural violation component + idiosyncratic noise, both lying in  $\perp S$ ). Quadratic-form concentration on the unspanned portion of  $\perp S$  (Vershynin 2018, Sec. 4.6) then yields the factor  $(1 - \kappa_{\text{eff}})$ .

The  $o_p(1)$  term follows from uniform concentration of quadratic forms and subspace perturbation bounds (sub-Gaussian tails + bounded 4th moments give  $O_p(\sqrt{\log T_0/T_0})$  rates).

**Regime-Specific Behavior** follows directly by substituting the realized  $\kappa_{\text{eff}}$ :  $s = 0 \Rightarrow \kappa_{\text{eff}} = 0$  (full recovery);  $0 < s < T_0 - r$  implies partial masking;  $s \geq T_0 - r \Rightarrow \kappa_{\text{eff}} \geq 1$  and full saturation.

### Remark 3.3 (Regularization and the Spurious Fit Trap)

The regularization analysis (Ridge, LASSO, Elastic Net) presented in the original appendix carries over qualitatively to the generalized setting: penalties modulate the effective dimension but cannot eliminate masking when  $\kappa_{\text{eff}} \geq 1$ . In the structured case the effective degrees of freedom become a function of both the penalty and the low-rank structure  $s$ , typically yielding residuals scaled by an unknown factor that still confounds structural bias with penalty bias. Cross-validation continues to drive effective penalties toward zero in saturated regimes, replicating the collapse observed under unstructured assumptions. Full details are analogous and omitted for brevity; the practical recommendation, donor curation remains first-order, continues to hold.

### Proposition 3.2 (Detection Boundary for Penalized Estimators)

Under Assumptions 2.1–2.5 and proportional asymptotics with  $\kappa > 0$ , there exist constants  $c_1, c_2 > 0$  such that:

- (i) Feasibility: If  $\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-3/2} \sqrt{(r + \log N)/T_0}$ , then a test based on the penalized residual norm detects violations with asymptotic power approaching 1.
- (ii) Impossibility: If  $\delta^2 < c_2 \sigma^2 (1 - \kappa)^{-3/2} \sqrt{(r + \log N)/T_0}$ , then no such test achieves power exceeding size +  $o(1)$ .

Moreover, the set of detectable violations is strictly smaller than for OLS (Theorem 4.2), as the  $\log N$  term inflates the boundary by a factor of  $\sqrt{\log N/r}$ , and the contamination exacerbates this via a higher exponent on  $(1-\kappa)^{-1}$  (specifically,  $-3/2$  vs.  $-1/2$  for OLS on the  $\delta^2$  scale, quantifying the gap as a factor of  $(1-\kappa)^{-1}$  worse for Lasso under selection uncertainty).

Even if the true counterfactual is sparse (only a few relevant donors exist), a high-dimensional pool of irrelevant donors ( $N_{\text{irrel}} \gg T_0$ ) generates a "shadow" sparse solution composed of irrelevant donors that fits the noise  $\epsilon_0$  better than the true donors fit the signal  $\lambda_0$ . To formalize, consider the LASSO solution path: the probability that a spurious support  $\hat{S}_{\text{spurious}}$  (from irrelevant donors) is selected before the true support  $\hat{S}_{\text{true}}$  is bounded below by

$$P(\text{spurious support selected first}) \geq 1 - \exp(-c\kappa \log N),$$

for some  $c > 0$ , as the maximum spurious correlation scales with  $\sqrt{\log N/T_0}$  (extreme value theory for sub-Gaussians; Vershynin, 2018), dominating the true signal when  $\kappa$  is moderate and  $N$  large. This occurs because, under contamination, the KKT threshold for irrelevant donors is violated with high probability via union bounds over  $N_{\text{irrel}}$ , creating misleading sparse approximations observationally equivalent to genuine ones under local alternatives (via contiguity as in Theorem 4.2).

#### Proof.

We derive the detection boundary for testing  $H_0: \delta^2 = 0$  (exact relevance) against  $H_1: \delta^2 > 0$  (relevance violation) using diagnostics based on the penalized residual norm from the LASSO estimator. The LASSO weights are

$$\hat{w} = \arg \min_w \|y_0 - Xw\|^2 + \lambda \|w\|_1,$$

with  $\lambda \asymp \sigma \sqrt{\log N/T_0}$  (theoretically optimal rate for sparse recovery under sub-Gaussian designs; Bickel et al., 2009, Theorem 7.2). The penalized residual is  $r = y_0 - X\hat{w}$ , and the normalized squared residual is  $R = \|r\|^2/T_0$ .

From Proposition 3.1 (generalized to structured irrelevance), under Assumptions 2.1–2.5, the pre-treatment outcome decomposes as  $y_0 = F\lambda_0 + \delta + \epsilon_0$ , where the relevance violation  $\delta$  (Definition 3.1) lies in the unattenuated subspace orthogonal to the donor span, with dimension  $T_0 - r - k_{\text{eff}}$  and  $\kappa = k_{\text{eff}}/(T_0 - r)$ .

We first establish the result under the dense regime (non-sparse true weights, many small entries in the population  $\lambda$ ), then extend to the sparse case as per the proposition statement.

**Dense Regime Setup.** For dense models (with true weights  $w^*$  satisfying  $\|w^*\|_0 \asymp N$ , entries  $O(1/\sqrt{N})$ ), LASSO does not achieve exact support recovery, as the irrepresentable condition fails with high probability (as shown in Section 3.3). The active set  $\hat{S}$  (selected donors) satisfies  $|\hat{S}| = O_p(N\kappa)$  under contamination, as spurious correlations with  $\delta$  and  $\epsilon_0$  dominate the KKT conditions (Wainwright, 2009, Theorem 2).

The penalized residual admits the decomposition

$$R = (1 - \kappa)\delta^2 + (1 - \kappa)\sigma^2 + b(\lambda) + o_p(1),$$

where  $b(\lambda)$  is the penalty-induced bias term. By oracle inequalities for LASSO under misspecification (Bühlmann & van de Geer, 2011, Section 6.4),  $b(\lambda) = O(\lambda |\hat{S}|/T_0) =$

$O(\sigma\kappa\log N)$ , since  $|\hat{S}| \asymp N\kappa$  and  $\lambda \asymp \sigma\sqrt{\log N/T_0}$ . The  $o_p(1)$  term follows from sub-Gaussian concentration on the residuals (Vershynin, 2018, Theorem 4.6.1, applied to quadratic forms).

Assume parameters  $(\kappa, \sigma, r)$  are known or consistently estimated (e.g.,  $\hat{\sigma}^2$  from donor spectral methods as in Remark 4.4;  $\hat{\kappa}$  from eigenvalue ratios, with limits per Remark 4.3). The test statistic is

$$T = \frac{R - \mathbb{E}[R | H_0]}{\hat{\sigma}_R},$$

where  $\hat{\sigma}_R$  estimates the standard deviation of  $R$  under  $H_0$ .

**(i) Asymptotic Feasibility (Dense Case).** Under  $H_1(\delta^2 > 0)$ , decompose  $y_0 = Xw^* + \delta + \epsilon_0$ . The LASSO residual satisfies

$$r = (I - H_\lambda)(\delta + \epsilon_0) + \text{bias from shrinkage},$$

where  $H_\lambda$  is the LASSO hat matrix (effective projector with degrees of freedom  $\text{df} \asymp |\hat{S}|$ ; Stein's unbiased risk estimate for LASSO). Since  $|\hat{S}| \asymp N\kappa$ , the effective subspace is saturated proportionally to  $\kappa$ , but LASSO shrinkage introduces bias on the order of  $\lambda \|\hat{w}\|_1/\sqrt{T_0} \asymp \sigma\sqrt{\kappa\log N}$  (from KKT stationarity conditions).

The unattenuated component of  $\delta$  lies in the noise subspace orthogonal to the relevant span, but spurious selections absorb a  $\kappa$ -fraction, yielding mean shift  $(1 - \kappa)\delta^2$ . The noise term is a quadratic form  $(1 - \kappa)\epsilon_0^\top(I - H_\lambda)\epsilon_0/T_0$ . Under sub-Gaussianity (Assumption 2.4),

$$\text{Var}(R | H_0) \asymp \sigma^2(r + \log N)(1 - \kappa)^{-1}/T_0,$$

as the  $\log N$  arises from model selection uncertainty in dense regimes (Wainwright, 2009, Corollary 1). By the CLT for penalized quadratic forms (under bounded higher moments and independence; extension of Vershynin, 2018, Section 4.6 via Lyapunov CLT),  $T \rightarrow_d \mathcal{N}(0,1)$  under  $H_0$ . Under  $H_1$ , the signal-to-noise ratio (SNR) is

$$\text{SNR} \asymp \frac{(1 - \kappa)\delta^2}{\sigma\sqrt{(r + \log N)(1 - \kappa)^{-1}/T_0}} = \frac{\delta^2}{\sigma(1 - \kappa)^{-3/2}\sqrt{(r + \log N)/T_0}}.$$

For  $\text{SNR} \rightarrow \infty$  (power  $\rightarrow 1$ ), require  $\delta^2 > c_1\sigma^2(1 - \kappa)^{-3/2}\sqrt{(r + \log N)/T_0}$ , absorbing constants (depending on sub-Gaussian parameters) into  $c_1$ .

**(ii) Impossibility (Dense Case).** Consider local alternatives  $\delta^2 \asymp \sigma^2(1 - \kappa)^{-3/2}\sqrt{(r + \log N)/T_0}$ . The shift in  $R$  matches the fluctuation order under  $H_0$ , including the  $\log N$  inflation from LASSO selection. The log-likelihood ratio (conditional on factors/loadings) for the penalized model is a quadratic form in sub-Gaussian residuals, modified by the subdifferential at the KKT point (non-smooth but satisfying local asymptotic normality conditions). Under density, this converges to a normal with finite variance under both hypotheses (by extensions of Le Cam's lemma to non-smooth estimators; van der Vaart, 2000, Theorem 6.4, adapted to  $\ell_1$ -penalized settings as in Bickel et al., 2009, Section 8). Mutual contiguity implies no test achieves power exceeding size +  $o(1)$ . For smaller  $\delta^2$ , contiguity holds a fortiori.

The boundary is strictly larger than for OLS (Theorem 4.2:  $\sigma^2(1 - \kappa)^{-1/2}\sqrt{r/T_0}$ ) by the factor  $\sqrt{(r + \log N)/r} \cdot (1 - \kappa)^{-1} \asymp \sqrt{\log N/r} \cdot (1 - \kappa)^{-1}$  (since  $\log N \gg r$  in proportional

asymptotics with  $N \asymp T_0$ ), confirming that LASSO's detectable set is a proper subset due to selection-induced variance inflation and bias confounding.

**Extension to Sparse True Model.** Assume the true model is sparse:  $w_S^*$  with support  $S$  of size  $s \ll N$ ,  $w_{S^c}^* = 0$ , and  $\min_j |w_j^*| > \lambda$  (beta-min condition for recovery). Even here, contamination induces spurious sparsity. The probability of spurious support selection first is derived as follows: The KKT for entry (irrelevant  $j$  enters if  $|X_j^T r_{-j}| \geq \lambda$ , where  $r_{-j} = \delta + \varepsilon_0 + \text{projection errors}$ ). Under sub-Gaussianity, the max over  $N_{\text{irrel}} \sim N$  of  $|X_j^T (\delta + \varepsilon_0)| / \sqrt{T_0} \sim \sqrt{(\log N / T_0)}$  with high probability (extreme value theory for Gaussians/sub-Gaussians; Vershynin, 2018, Proposition 2.1.2). Since  $\lambda \sim \sigma \sqrt{(\log N / T_0)}$ , the event  $\max \text{spurious corr} > \lambda$  has  $P \geq 1 - \exp(-c \log N)$  (union bound over  $N$ , tail probability  $\exp(-t^2)$  for  $t \sim \text{constant}$ ). For  $\kappa < 1$  (structured), the effective "independent looks" are reduced to  $\kappa T_0$  dimensions, but the dictionary density ( $N$  vectors in low-rank space) allows covering numbers scaling with  $N$ , preserving  $\log N$  extremes (random projection theorem; Johnson-Lindenstrauss). When true signals satisfy  $w_{0i} \lesssim \sqrt{(\log N / T_0)}$ , spurious exceed true; stronger signals recover, but widens twilight zone. Contiguity for local  $\delta$  (scaled by boundary) shows equivalence. Finite-sample via concentration.  $\square$

## Appendix A.4 (proofs of Section 4)

### Theorem 4.1 (Impossibility of Testing Exact Relevance Without Separation).

Under Assumptions 2.1–2.5 and proportional asymptotics with  $N_{\text{rel}}/T_0 \rightarrow \nu \in [0, \infty)$  and  $\kappa \in [0, 1)$  (avoiding the spurious fit trap), consider testing

$$H_0: \delta = 0 \text{ versus } H_{1,T}: \delta_T = c\sigma_\epsilon \sqrt{\frac{r}{T_0}}, c > 0$$

Without imposing a minimum separation condition on  $\delta$ , no projection-based test based solely on pre-treatment outcomes  $\{y_0, X\}$  can simultaneously:

1. Control asymptotic size at level  $\alpha$  under  $H_0$ , and
2. Achieve power exceeding  $\alpha$  against the local alternatives  $H_{1,T}$  as  $T_0 \rightarrow \infty$ .

The constant  $c$  in the local alternative  $\delta_T = c\sigma_\epsilon \sqrt{r/T_0}$  depends on the factor covariance structure  $\Sigma_F$  through Definition 3.1's metric, but the impossibility result holds for any fixed  $\Sigma_F \succ 0$ .

While our asymptotic framework treats the factor dimension  $r$  as fixed (consistent with Assumptions 2.1–2.5), we explicitly track the dependence on  $r$  in finite-sample expressions and detection rates (e.g.,  $O(\sqrt{r/T_0})$ ) to illustrate how diagnostic power scales with the complexity of the latent confounding structure. This does not alter the asymptotic rates, where such terms are absorbed into constants, but provides practical calibration for applications where  $r$  may be moderate relative to  $T_0$ .

#### Proof.

We establish contiguity of the null and local alternative distributions using Le Cam's lemma. Fix the factor path  $F$  and donor loadings  $\Lambda$  (this conditioning is valid under Assumption 2.5, which treats factors as non-random or conditions on their realization).

Under  $H_0: \delta = 0$ , we have  $\mathbb{E}[y_0] = F\lambda_0$  with  $\lambda_0 \in \text{span}(\Lambda)$ .

Under  $H_{1,T}: \delta_T = c\sigma_\epsilon \sqrt{r/T_0}$ , we have  $\mathbb{E}[y_0] = F\lambda_0 + \mu_T$  where  $\|\mu_T\|^2 = \delta_T^2 T_0 = c^2 \sigma_\epsilon^2 r$ .

The pre-treatment outcome  $y_0$  satisfies:

$$y_0 = \mathbb{E}[y_0] + \epsilon_0$$

where  $\epsilon_0 \sim N(0, \sigma_\epsilon^2 I_{T_0})$  under Assumption 2.4 (sub-Gaussian implies Gaussian for the CLT argument; for general sub-Gaussian, use characteristic function arguments).

The log-likelihood ratio is:

$$\ell_T = \log \frac{p_{H_{1,T}}(y_0)}{p_{H_0}(y_0)} = \frac{1}{\sigma_\epsilon^2} \left[ \mu_T'(y_0 - \mathbb{E}[y_0]) - \frac{1}{2} \|\mu_T\|^2 \right]$$

Substituting  $y_0 = F\lambda_0 + \epsilon_0$  under  $H_0$ :

$$\ell_T = \frac{1}{\sigma_\epsilon^2} \mu_T' \epsilon_0 - \frac{1}{2\sigma_\epsilon^2} \|\mu_T\|^2$$

The first term has mean zero and variance:

$$\text{Var}(\mu'_T \epsilon_0) = \sigma_\epsilon^2 \|\mu_T\|^2 = \sigma_\epsilon^2 \cdot c^2 \sigma_\epsilon^2 r = c^2 \sigma_\epsilon^4 r$$

Therefore:

$$\ell_T = \frac{\mu'_T \epsilon_0}{\sigma_\epsilon^2} - \frac{c^2 \sigma_\epsilon^2 r}{2\sigma_\epsilon^2} \sim N\left(-\frac{c^2 r}{2}, c^2 r\right) \text{ under } H_0$$

By standard weak convergence results (van der Vaart 2000, Theorem 6.4), the likelihood ratio converges in distribution to a non-degenerate normal random variable under both hypotheses. Specifically:

$$\begin{aligned} \ell_T &\xrightarrow{d} N\left(-\frac{c^2 r}{2}, c^2 r\right) \text{ under } H_0 \\ \ell_T &\xrightarrow{d} N\left(+\frac{c^2 r}{2}, c^2 r\right) \text{ under } H_{1,T} \end{aligned}$$

These distributions have positive Hellinger affinity:

$$H^2(P_0, P_{1,T}) = 1 - \exp\left(-\frac{c^4 r^2}{8c^2 r}\right) = 1 - \exp\left(-\frac{c^2 r}{8}\right) < 1$$

By Le Cam's third lemma (van der Vaart 2000, Lemma 6.4), the sequences  $\{P_0^{(T)}\}$  and  $\{P_{1,T}^{(T)}\}$  are mutually contiguous. Contiguity implies:

1. Any test with asymptotic size  $\alpha$  under  $H_0$  has asymptotic power at most  $\alpha + o(1)$  under  $H_{1,T}$ .
2. No sequence of tests can distinguish  $H_0$  from  $H_{1,T}$  with error probabilities vanishing to zero.

This establishes the impossibility of consistent testing without imposing a minimum separation exceeding  $\sigma_\epsilon \sqrt{r/T_0}$ .  $\square$

#### **Theorem 4.2 (Detection Boundary for Projection-Based Diagnostics)**

Under Assumptions 2.1–2.5, suppose  $0 < \kappa < 1$ . There exist constants  $c_1, c_2 > 0$ , depending only on distributional parameters and  $r$ , such that:

**(i-a) Asymptotic Feasibility.** If

$$\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r}{T_0}},$$

then there exists a test based on the projection residual norm whose asymptotic power converges to one while controlling size at any fixed level.

**(i-b) Finite-Sample Feasibility.** If

$$\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r + \log T_0}{T_0}},$$

for some constant  $c_1 > 0$  and fixed  $\alpha > 0$ , then violations exceeding this are detectable with probability at least  $1 - \alpha$ .

**(ii) Impossibility.** If

$$\delta^2 < c_2 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r}{T_0}},$$

then no test measurable with respect to the projection residuals attains asymptotic power exceeding size plus  $o(1)$ .

Moreover, as  $\kappa \rightarrow 1$  (saturation of the outcome space), the boundary diverges for fixed  $\delta^2$ , implying that relevance violations become asymptotically undetectable via projection-based residual diagnostics.

**Proof.**

We establish the detection boundary for testing  $H_0: \delta^2 = 0$  (exact relevance) against  $H_1: \delta^2 > 0$  (relevance violation) using diagnostics based on the projection residual norm. The key statistic is the normalized squared residual  $R = \|r\|^2 / T_0$ , where  $r = y_0 - P_X y_0$  and  $P_X$  is the orthogonal projection onto the donor span  $\text{span}(X)$ .

From Proposition 3.1, under Assumptions 2.1–2.5,

$$R = (1 - \kappa)\delta^2 + (1 - \kappa)\sigma^2 + o_p(1),$$

for  $0 < \kappa < 1$ .

Assume parameters  $(\kappa, \sigma, r)$  are known or consistently estimated (e.g.,  $\hat{\delta}^2$  from donor spectral methods as in Remark 4.4;  $\hat{\kappa}$  from eigenvalue ratios, with limits per Remark 4.3). The test statistic is

$$T = \frac{R - \mathbb{E}[R | H_0]}{\hat{\sigma}_R},$$

where  $\hat{\sigma}_R$  estimates the standard deviation of  $R$  under  $H_0$ .

**(i-a) Asymptotic Feasibility.** Under  $H_1 (\delta^2 > 0)$ , decompose  $y_0 = Xw^* + \delta + \epsilon_0$ , where  $w^*$  satisfies the relevance condition. The residual is

$$r = (I - P_X)(\delta + \epsilon_0).$$

The unattenuated component of  $\delta$  lies in the noise subspace orthogonal to the donor span, yielding a structural mean shift of  $(1 - \kappa)\delta^2$ . The noise term is the quadratic form  $(1 - \kappa)\epsilon_0^\top (I - P_X)\epsilon_0 / T_0$ . Under sub-Gaussianity (Assumption 2.4), the variance scales with the effective dimension of the noise subspace:

$$\text{Var}(R | H_0) \asymp \sigma^2 r (1 - \kappa) / T_0,$$

where  $r = \text{rank}_{\text{eff}}((I - P_X)\Sigma_\epsilon)$ , the effective noise dimension in the orthogonal subspace (or, formally,  $\text{tr}(((I - P_X)\Sigma_\epsilon)^2) \asymp \sigma^2 r (1 - \kappa)$ ). By the CLT for quadratic forms (under independence and bounded moments; under Assumption 2.5 ensuring weak temporal dependence or conditional independence given factors, standard quadratic-form CLTs apply),  $T \rightarrow_d \mathcal{N}(0, 1)$  under  $H_0$ . Under  $H_1$ , the signal-to-noise ratio (SNR) is:

$$\text{SNR} \asymp \frac{(1 - \kappa)\delta^2}{\sigma\sqrt{r(1 - \kappa)/T_0}} = \frac{\delta^2(1 - \kappa)^{1/2}}{\sigma\sqrt{r/T_0}}.$$

For  $\text{SNR} \rightarrow \infty$ , require:

$$\delta^2 > c_1 \sigma^2 (1 - \kappa)^{-1/2} \sqrt{\frac{r}{T_0}},$$

absorbing constants into  $c_1$ .

**(i-b) Finite-Sample Feasibility.** Sub-Gaussian concentration on quadratic forms (Vershynin, 2018, Theorem 4.6.1) provides high-probability bounds. The deviation of the quadratic form from its mean scales with the square root of the variance plus a tail term involving  $\log T_0$ . With probability at least  $1 - \alpha$ , the noise fluctuation is bounded by  $O(\sigma^2(r + \log T_0)(1 - \kappa)/T_0)$ . Solving for the signal dominance yields the log-adjusted threshold.

**(ii) Impossibility.** Consider local alternatives  $\delta^2 \asymp \sigma^2(1 - \kappa)^{-1/2}\sqrt{r/T_0}$ . The shift in  $R$  matches the fluctuation order under  $H_0$ . The log-likelihood ratio (conditional on factors/loadings) is a quadratic form in sub-Gaussian residuals, converging to a normal with finite variance under both hypotheses (Le Cam's lemma; van der Vaart, 2000, Theorem 6.4). The experiment reduces to testing the mean of an asymptotically normal statistic with variance  $\asymp \sigma^2 r(1 - \kappa)/T_0$ , which makes Le Cam immediate. Mutual contiguity implies no test achieves power exceeding size +  $o(1)$ . For smaller  $\delta^2$ , contiguity holds a fortiori. As  $\kappa \rightarrow 1$ , the boundary diverges, as the unattenuated subspace shrinks to zero dimension. Finite-sample analogs follow from concentration.  $\square$

#### Convergence of Donor-Based Variance Estimator.

Let  $\hat{\sigma}_X^2$  be the variance estimator derived from the trailing  $N - \hat{r}$  eigenvalues of the donor Gram matrix  $\hat{\Sigma}_X = X'X/T_0$ , where  $\hat{r}$  is consistent for the factor number (e.g., Ahn & Horenstein, 2013). Under Assumptions 2.1–2.3 (homoskedastic errors across units), the donor matrix  $X$  follows a spiked covariance model. By the Marchenko-Pastur law and Weyl's inequality (Vershynin 2018, Thm 4.6.1), the eigenvalues  $\lambda_j(\hat{\Sigma}_X)$  for  $j > r$  concentrate around  $\sigma_\epsilon^2$ .

Consequently,  $\hat{\sigma}_X^2 \xrightarrow{p} \sigma_\epsilon^2$ . Unlike the residual-based estimator  $\hat{\sigma}_{resid}^2 = \|r\|^2/T_0$ , which converges to  $(1 - \alpha)\sigma_\epsilon^2 + \delta^2(1 - \kappa)$  (Proposition 3.1), the donor-based estimator is invariant to the projection geometry  $P_X$ . Thus, the test statistic  $\tau$  constructed using  $\hat{\sigma}_X^2$  avoids the cancellation of the  $(1 - \kappa)$  term, ensuring that the power bounds in Theorem 4.2 hold.

#### 4.3 Weak Conditioning and Non-Regular Inference

We now turn to conditioning. Even when relevance holds ( $\delta^2 = 0$ ), ill-conditioned factor or donor spaces lead to unstable estimation and nonstandard inference. Let

$$\lambda_{\min}(G_T) = \min \left\{ \lambda_{\min} \left( \frac{F^\top F}{T_0} \right), \lambda_{\min} \left( \frac{\Lambda^\top \Lambda}{N} \right) \right\}$$

denote the minimum eigenvalue of the Gram matrices governing temporal (factor) and cross-sectional (donor) conditioning.

#### Theorem 4.3 (Non-Regular Asymptotics under Weak Conditioning)

Suppose relevance holds ( $\lambda_0 \in \text{span}(\Lambda)$ ) and the minimum eigenvalue satisfies  $\lambda_{\min} \rightarrow 0$  at rate  $T_0^{-\gamma}$  for  $\gamma \in (0,1)$ . Consider the least-squares weight estimator  $\hat{w}$ . Then:

**(I) Variance divergence:** The estimation variance satisfies  $\text{Var}(\hat{w}) \sim \Omega(T_0^{\gamma-1})$ , so  $\sqrt{T_0}$ -consistency fails:  $\hat{w} - w = O_p(T_0^{(\gamma-1)/2})$ .

**(II) Rescaled convergence:** The appropriately rescaled estimator  $T_0^{(1-\gamma)/2}(\hat{w} - w)$  converges to a non-degenerate limit whose distribution depends on the eigenvector structure of the weak direction.

**(III) Non-Gaussian limits:** When weak conditioning arises from near-singular donor loadings ( $\lambda_{\min}(\Lambda\Lambda') \rightarrow 0$ ), the asymptotic distribution is non-Gaussian (specifically, a ratio of random variables analogous to weak-IV asymptotics).

**Simplified Gaussian Case.** Under Gaussian factors and errors ( $f_t \sim N(0, I_r)$ ,  $\epsilon_{it} \sim N(0, \sigma^2)$ ), assume the weak conditioning is driven by a single vanishing eigenvalue  $\lambda_{\min}(\hat{G}) \sim cT_0^{-\gamma}$  along eigenvector  $v_N$ . The rescaled error along this direction is

$$T_0^{(1-\gamma)/2}[(\hat{w} - w) \cdot v_N] \rightarrow_d \frac{Z}{\xi},$$

where  $Z \sim N(0, \sigma^2)$  is the projected noise, and  $\xi \sim \chi_1^2/c$  (rescaled chi-squared from the eigenvalue limit). This non-standard ratio invalidates Gaussian inference, with tails heavier than normal. Explicit characterization for general cases requires additional regularity on the factor process and is left to future work. For practical inference, weak-conditioning-robust methods (analogous to Anderson-Rubin in IV) are required.

### Proof.

We prove the theorem under Assumptions 2.1–2.5. Since relevance holds, there exists a population weight vector  $w \in \mathbb{R}^N$  such that  $\mathbb{E}[y_0 \mid F, \Lambda] = Xw$ , where  $y_0$  is the  $T_0 \times 1$  pre-treatment vector for the treated unit, and  $X$  is the  $T_0 \times N$  donor matrix. The least-squares estimator is

$$\hat{w} = (X'X)^{-1}X'y_0.$$

The estimation error is

$$\hat{w} - w = (X'X)^{-1}X'(y_0 - Xw) = (X'X)^{-1}X'\epsilon_0,$$

where  $\epsilon_0$  is the  $T_0 \times 1$  vector of idiosyncratic errors for the treated unit. By Assumption 2.2 (strict exogeneity),  $\epsilon_0$  is independent of  $X$  (and thus of the Gram matrix  $X'X$ ) with  $\mathbb{E}[\epsilon_{0t}] = 0$  and  $\text{Var}(\epsilon_{0t}) = \sigma^2$ .

Let  $\hat{G} = X'X/T_0$  be the sample Gram matrix, with eigenvalues  $\lambda_1(\hat{G}) \geq \dots \geq \lambda_N(\hat{G})$  and corresponding orthonormal eigenvectors  $v_1, \dots, v_N$ . Weak conditioning is formalized as

$$\lambda_{\min}(\hat{G}) = \lambda_N(\hat{G}) \rightarrow_p 0 \text{ at rate } \lambda_{\min}(\hat{G}) \sim cT_0^{-\gamma}, \gamma \in (0,1),$$

for some constant  $c > 0$ .

### Step 1: Variance Inflation Along the Weak Direction

Let  $v$  be the eigenvector associated with  $\lambda_{\min}$ , normalized so  $\|v\| = 1$ . The projection of the estimation error onto this direction is

$$v'(\hat{w} - w) = v'(X'X)^{-1}X'\epsilon_0 = \frac{1}{T_0}v'\hat{G}^{-1}(X'\epsilon_0).$$

Because  $\hat{G}v = \lambda_{\min}v$ , we have  $\hat{G}^{-1}v = \lambda_{\min}^{-1}v$ , and therefore

$$v'\hat{G}^{-1} = \lambda_{\min}^{-1}v'.$$

Substituting yields

$$v'(\hat{w} - w) = \frac{1}{T_0\lambda_{\min}}v'X'\epsilon_0 = \frac{1}{T_0\lambda_{\min}}\sum_{t=1}^{T_0}z_t\epsilon_{0t},$$

where  $z_t = X_t'v$  and  $X_t'$  denotes the  $t$ -th row of  $X$ .

The variance of the numerator is

$$\text{Var}\left(\sum_{t=1}^{T_0}z_t\epsilon_{0t}\right) = \sigma^2\sum_{t=1}^{T_0}z_t^2 = \sigma^2\|Xv\|^2 = \sigma^2v'X'Xv = \sigma^2T_0v'\hat{G}v = \sigma^2T_0\lambda_{\min},$$

where the second equality uses independence of  $\epsilon_0$  and  $X$ . Therefore,

$$\text{Var}(v'(\hat{w} - w)) = \left(\frac{1}{T_0\lambda_{\min}}\right)^2 \cdot \sigma^2T_0\lambda_{\min} = \frac{\sigma^2}{T_0\lambda_{\min}}.$$

Substituting the rate  $\lambda_{\min} \sim cT_0^{-\gamma}$  gives

$$\text{Var}(v'(\hat{w} - w)) \sim \frac{\sigma^2}{cT_0^{\gamma-1}} \sim \Omega(T_0^{\gamma-1}).$$

This establishes the variance divergence claimed in part (i):  $\text{Var}(\hat{w}) = \Omega_p(T_0^{\gamma-1})$  (dominated by the weak direction), and consequently

$$\hat{w} - w = O_p(T_0^{(\gamma-1)/2}).$$

The standard  $\sqrt{T_0}$ -consistency rate fails whenever  $\gamma > 0$ .

## Step 2: Rescaled Convergence

Multiplying by the normalizing sequence  $T_0^{(1-\gamma)/2}$  yields

$$T_0^{(1-\gamma)/2}(\hat{w} - w) = O_p(1)$$

along the weak direction. The limit is non-degenerate because the numerator  $\sum_{t=1}^{T_0}z_t\epsilon_{0t}$  is a sum of independent terms with positive variance of order  $T_0\lambda_{\min}$ , which exactly offsets the denominator's order.

## Step 3: Non-Gaussianity

The limiting distribution of the rescaled error is non-Gaussian in general. This occurs because:

- The weak eigenvector  $v$  is itself a random function of  $X$  (and thus of the noise matrix  $E$  in  $X = F\Lambda' + E$ ).
- The projected process  $z_t = X_t'v$  is therefore stochastically dependent on the direction of weakness.

The limit typically takes the form of a ratio of quadratic forms (or a normal divided by a random denominator), analogous to the weak-instrument asymptotics of Staiger & Stock (1997). Explicit derivation of the limiting distribution requires additional structure on the joint distribution of the factor process and the idiosyncratic errors (e.g., Gaussian factors or specific mixing conditions), which we do not impose. We therefore leave the full characterization for future work.  $\square$

## Appendix A.5 (proofs of Section 5)

### Theorem 5.1 (Existence and Point Identification of Counterfactuals)

Under the factor structure (Assumption 2.3) and bounded moments (Assumption 2.4), a latent counterfactual representation exists and is uniquely identified if and only if the relevance condition  $\lambda_0 \in \text{span}(\Lambda)$  holds.

**Moreover:**

**(i) Non-Existence (Irreducible Error):** If relevance fails ( $\lambda_0 \notin \text{span}(\Lambda)$ ), then for any measurable estimator  $\hat{Y}_{0t}(0)$  based on donor histories  $\{X_i\}_{i=1}^N$ , the conditional bias is bounded below:

$$\liminf_{T_0 \rightarrow \infty} \mathbb{E}[\hat{Y}_{0t}(0) | f_t] - Y_{0t}(0) \geq c \cdot \delta \cdot \|f_t\|$$

where  $\delta = \text{dist}(\lambda_0, \text{span}(\Lambda))$  and  $c > 0$  depends on the factor covariance structure.

Consequently, the Mean Squared Error (MSE) is bounded away from zero regardless of sample size.

**(ii) Uniqueness:** If relevance holds, the counterfactual outcome path  $\{Y_{0t}(0)\}_{t>T_0}$  is uniquely determined (in expectation) by:

$$\mathbb{E}[Y_{0t}(0) | f_t] = \lambda_0' f_t = w_0' X_t$$

for any weight vector  $w_0$  satisfying  $\lambda_0 = \Lambda' w_0$ .

**Proof.**

**Part (i):** Under the assumption that factors are generated by a stable process (Assumption 2.5), the information in the donor outcomes  $X_t = \Lambda f_t + \epsilon_t$  is restricted to the subspace spanned by  $\Lambda$ . Decompose the treated unit's loading into in-span and orthogonal components:

$$\lambda_0 = P_\Lambda \lambda_0 + (I - P_\Lambda) \lambda_0 = \lambda_0^\parallel + \lambda_0^\perp$$

The outcome is:

$$Y_{0t}(0) = (\lambda_0^\parallel)' f_t + (\lambda_0^\perp)' f_t + \epsilon_{0t}$$

The first component is recoverable from donors since  $\lambda_0^\parallel \in \text{span}(\Lambda)$ . The second component involves  $\lambda_0^\perp$ , which is orthogonal to the column space of  $\Lambda'$ . Since donor outcomes  $X_t$  depend only on  $\Lambda f_t$ , they contain no information about the projection of  $f_t$  onto  $\lambda_0^\perp$  (assuming the factor covariance  $\Sigma_F$  is positive definite).

By sufficiency of linear projections under Gaussianity (or applying the completeness result of Newey & Powell, 2003, to rule out non-linear recovery), the expected squared error of any estimator satisfies:

$$\mathbb{E}[(\hat{Y}_{0t}(0) - Y_{0t}(0))^2] \geq \mathbb{E}[(\lambda_0^\perp)' f_t]^2 = \|\lambda_0^\perp\|^2 \cdot \mathbb{E}[f_t' f_t]$$

Noting  $\|\lambda_0^\perp\| = \delta$ , this implies a lower bound on the error scaling with  $\delta$ .

**Part (ii):** Follows from the linearity of the factor model. Any two weight vectors  $w, w'$  satisfying  $\Lambda' w = \Lambda' w' = \lambda_0$  produce identical expected outcomes:

$$w' X_t = w' \Lambda f_t = (w' \Lambda) f_t = \lambda_0' f_t = (w'' \Lambda) f_t = w'' X_t$$

The difference  $w - w'$  lies in  $\text{null}(\Lambda')$ , which implies  $(w - w')'X_t = 0$  for all realizations of  $X_t$  generated by the factors. Thus, the systematic counterfactual component is invariant to the choice of weights.  $\square$

### Theorem 5.2 (Conditioning and Regular Inference)

Suppose relevance holds ( $\lambda_0 \in \text{span}(\Lambda)$ ) so that the latent counterfactual representation is uniquely identified. Consider projection-based estimators obtained as solutions to:

$$\hat{w} = \arg \min_{w \in \mathbb{R}^N} \|y_0 - Xw\|_2^2$$

or regularized variants (Ridge, LASSO, Elastic Net).

**(i) Regular asymptotics:** If persistent conditioning holds (Assumption 2.5), then under standard moment conditions (Assumptions 2.2–2.3), the estimator admits a  $\sqrt{T_0}$ -consistent and asymptotically normal representation:

$$\sqrt{T_0}(\hat{w} - w_0) \xrightarrow{d} N(0, \sigma_\epsilon^2 \Sigma_X^{-1})$$

where  $\Sigma_X = \text{plim}(X'X/T_0)$ , and inference based on asymptotic normality is valid.

**(ii) Weak identification:** If the minimum eigenvalue converges to zero at rate:

$$\lambda_{\min}\left(\frac{X'X}{T_0}\right) \asymp T_0^{-\gamma}, \gamma \in (0,1)$$

then no sequence of projection-based estimators achieves uniform  $\sqrt{T_0}$ -consistency. Specifically:

- **Variance Inflation:**  $\mathbb{E} \|\hat{w} - w_0\|^2 = \Omega(T_0^{\gamma-1})$ . While the estimator remains consistent (variance  $\rightarrow 0$ ), it converges strictly slower than the standard  $T_0^{-1}$  rate. The non-Gaussian limiting distribution arises because the projection  $v_{\min}X'\epsilon_0$  onto the weak direction has variance  $O(T_0\xi_{\min}) = O(1)$  when  $\xi_{\min} \asymp T_0^{-1}$ , while the normalization  $(X'X)^{-1}$  amplifies this by  $T_0^\gamma$ . This creates a ratio of random quadratic forms similar to weak-IV asymptotics (see Staiger-Stock 1997, Theorem 1). Fully characterizing this distribution is beyond our scope.
- **Rescaled Convergence:** The appropriately rescaled estimator  $T_0^{(1-\gamma)/2}(\hat{w} - w_0)$  converges to a non-degenerate limit.
- **Non-Gaussian Limits:** Asymptotic distributions are generically non-normal, defined by ratios of random variables (noise projected onto weak eigenvectors), invalidating standard t-tests and confidence intervals.

**(iii) Complete failure:** If  $\lambda_{\min}(X'X/T_0) \rightarrow 0$  at rate  $T_0^{-1}$  (e.g., when  $N/T_0 \rightarrow \infty$  or  $\kappa > 1$ ), then:

$$\|\hat{w}\| = O_p(T_0) \text{ and } \|\hat{w} - w_0\| = O_p(T_0)$$

The weights diverge and counterfactual estimates are numerically unstable and uninformative regardless of relevance.

**Proof.**

**(Part i):** Under persistent conditioning,  $X'X/T_0 \xrightarrow{p} \Sigma_X$  with  $\lambda_{\min}(\Sigma_X) > 0$ . By the CLT (Assumption 2.2),  $X'\epsilon_0/\sqrt{T_0} \xrightarrow{d} N(0, \sigma_\epsilon^2 \Sigma_X)$ . Slutsky's theorem yields the standard result:

$$\sqrt{T_0}(\hat{w} - w_0) = \left( \frac{X'X}{T_0} \right)^{-1} \frac{X'\epsilon_0}{\sqrt{T_0}} \xrightarrow{d} \Sigma_X^{-1} N(0, \sigma_\epsilon^2 \Sigma_X) = N(0, \sigma_\epsilon^2 \Sigma_X^{-1})$$

**(Part ii):** Follows from Theorem 4.3. When  $\lambda_{\min} \asymp T_0^{-\gamma}$ , the inverse Gram matrix scales as  $\Omega(T_0^\gamma)$ . The error term along the weak direction behaves as:

$$[\hat{w} - w_0]_{\text{weak}} \approx T_0^\gamma \cdot O_p(T_0^{-\frac{1}{2}}) = O_p(T_0^{\gamma-\frac{1}{2}})$$

The variance scales as squared error:  $T_0^{2\gamma-1}$ .

Comparing this to the standard rate  $T_0^{-1}$ :

$$\text{Rate Ratio} = \frac{T_0^{2\gamma-1}}{T_0^{-1}} = T_0^{2\gamma} \rightarrow \infty$$

Since the variance decays slower than  $1/T_0$ , multiplying by  $\sqrt{T_0}$  (as in a t-statistic) causes the distribution to diverge, invalidating standard inference.

**(Part iii):** When  $\lambda_{\min}(X'X/T_0) = O(T_0^{-1})$ , the smallest eigenvalue of the unscaled matrix  $X'X$  is  $O(1)$ . Thus,  $(X'X)^{-1} = O(1)$ . The projection  $X'y_0$  scales as  $O_p(T_0)$  (sum of  $T_0$  terms). Therefore:

$$\hat{w} = (X'X)^{-1}X'y_0 \approx O(1) \cdot O_p(T_0) = O_p(T_0)$$

The weights explode as  $T_0$  increases.  $\square$

### Proposition 5.3 (Spurious Fits under Simplex Constraints).

Under Assumptions 2.1–2.5 and the contamination model, consider the synthetic control estimator

$$\hat{w}_{sc} = \arg \min_w \|Y_0 - Xw\|^2 \text{ s.t. } w \geq 0, 1^\top w = 1.$$

With probability approaching 1 as  $\text{rank}(X_{irrel})/T_0 \rightarrow 1$ , if the relevance violation component  $F\delta$  lies approximately in the convex hull of the irrelevant donor outcomes  $X_{irrel}$ , then the normalized squared residuals attenuate proportionally to an effective contamination ratio  $\kappa_{eff} < \kappa$ , where  $\kappa_{eff}$  scales with  $(\log N_{irrel})/T_0$  rather than  $\text{rank}(X_{irrel})/T_0$ . Otherwise, if  $F\delta$  lies outside the scaled convex hull, residuals remain bounded away from zero.

### Proof

Decompose the pre-treatment treated outcome as

$$Y_0 = S_p + F\delta + \epsilon_0,$$

where  $S_p = FP\lambda_0$  lies in the relevant factor span,  $F\delta$  is the structural violation in the orthogonal complement (with population margin  $m = \|F\delta\|^2/T_0$ ), and  $\epsilon_0$  is sub-Gaussian noise (we set  $\epsilon_0 \approx 0$  for the leading term).

Partition  $X = [X_{rel}, X_{irrel}]$  and  $\hat{w}_{sc} = (w_{rel}, w_{irrel})$ . Let  $\alpha = 1^\top w_{irrel}$  (so  $1^\top w_{rel} = 1 - \alpha$ ) and  $v = w_{irrel}/\alpha$  (with  $v \geq 0, 1^\top v = 1$ ).

The fitted synthetic control is approximately

$$X\hat{w}_{sc} \approx (1 - \alpha)S_p + \alpha(X_{irrel}v).$$

The squared residual then becomes

$$\|Y_0 - X\hat{w}_{sc}\|^2 \approx \alpha^2 \|S_p\|^2 + \|\alpha(X_{irrel}v) - F\delta\|^2.$$

Define  $d(\alpha) = \min_{v \geq 0, 1^\top v = 1} \|\alpha(X_{irrel}v) - F\delta\|$ . The objective reduces to minimizing

$$f(\alpha) = \alpha^2 \|S_p\|^2 + d(\alpha)^2, \alpha \in [0, 1].$$

Under sub-Gaussianity, the convex hull  $\text{conv}(X_{irrel})$  behaves like a random polytope. The maximum projection of the hull in the direction of  $u_\delta = F\delta/\|F\delta\|$  satisfies

$$C := \max_v \langle X_{irrel}v, u_\delta \rangle \approx \sigma \sqrt{2 \log N_{irrel}} \text{ w.h.p.}$$

For moderate violations where  $\alpha C < \|F\delta\|$ , we have  $d(\alpha) \approx \|F\delta\| - \alpha C$ . Substituting and differentiating  $f(\alpha)$  yields the optimal weight

$$\alpha^* = \frac{\|F\delta\| \cdot C}{\|S_p\|^2 + C^2}.$$

The minimized residual is then

$$\|Y_0 - X\hat{w}_{sc}\|^2 \approx \|F\delta\|^2 \cdot \frac{\|S_p\|^2}{\|S_p\|^2 + C^2}.$$

Normalizing by  $T_0$ , the squared relevance margin  $m$  is attenuated by the factor

$$\frac{\|S_p\|^2}{\|S_p\|^2 + C^2} \approx \frac{T_0}{T_0 + 2\sigma^2 \log N_{irrel}}.$$

Thus, the effective contamination ratio becomes

$$\kappa_{eff} \approx \frac{2\sigma^2 \log N_{irrel}}{T_0} \ll \kappa = \frac{\text{rank}(X_{irrel})}{T_0},$$

for typical panel data where  $\log N_{irrel} \ll T_0$  and  $\text{rank}(X_{irrel}) \approx \min(N_{irrel}, T_0)$ .

If  $\|F\delta\| > C$  (large violation or poor alignment),  $d(\alpha) > 0$  remains bounded away from zero, so residuals do not vanish and violations are more easily detected than under OLS (which can achieve near-zero residuals using negative weights).  $\square$

**Remark.** Full uniform concentration over directions uses  $\varepsilon$ -nets and Vershynin (2018, Theorem 4.6.1).