Journal of Machine Learning Research 26 (2025) 1-S33 Submitted 1/23; Revised 2/25; Published 6/25

On Model Identification and Out-of-Sample Prediction of PCR with Applications to Synthetic Controls

Anish Agarwal AA5194@COLUMBIA.EDU
Department of Industrial Engineering & Operations Research
Columbia University

Devavrat Shah DEVAVRAT@MIT.EDU
Department of Electrical Engineering & Computer Science
Massachusetts Institute of Technology

Dennis Shen DENNIS.SHEN@MARSHALL.USC.EDU
Department of Data Sciences & Operations
University of Southern California

Editor: Ilya Shpitser

Abstract

We analyze principal component regression (PCR) in a high-dimensional error-in-variables setting with fixed design. Under suitable conditions, we show that PCR consistently identifies the unique model with minimum ℓ_2-norm. These results enable us to establish non-asymptotic out-of-sample prediction guarantees that improve upon the best known rates. In the course of our analysis, we introduce a natural linear algebraic condition between the in- and out-of-sample covariates, which allows us to avoid distributional assumptions for out-of-sample predictions. Our simulations illustrate the importance of this condition for generalization, even under covariate shifts. Accordingly, we construct a hypothesis test to check when this condition holds in practice. As a byproduct, our results also lead to novel results for the synthetic controls literature, a leading approach for policy evaluation. To the best of our knowledge, our prediction guarantees for the fixed design setting have been elusive in both the high-dimensional error-in-variables and synthetic controls literatures.

Keywords: error-in-variables, fixed design, high-dimensional, covariate shift, missing data

1. Introduction

We consider error-in-variables regression in a high-dimensional setting with fixed design. Formally, we observe a labeled dataset of size n, denoted as $\{(y_i, \boldsymbol{z}_i) : i \leq n\}$. Here, $y_i \in \mathbb{R}$ is the response variable and $\boldsymbol{z}_i \in \mathbb{R}^p$ is the observed covariate. For any $i \geq 1$, we posit that

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle + \varepsilon_i, \tag{1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the unknown model parameter, $\boldsymbol{x}_i \in \mathbb{R}^p$ is a fixed covariate, and $\varepsilon_i \in \mathbb{R}$ is the response noise. Unlike traditional settings where $\boldsymbol{z}_i = \boldsymbol{x}_i$, the error-in-variables (EiV) setting reveals a corrupted version of the covariate $\boldsymbol{x}_i$. Precisely, for any $i \geq 1$, let

$$\boldsymbol{z}_i = (\boldsymbol{x}_i + \boldsymbol{w}_i) \circ \boldsymbol{\pi}_i, \tag{2}$$

where $\boldsymbol{w}_i \in \mathbb{R}^p$ is the covariate measurement noise, $\boldsymbol{\pi}_i \in \{1, \mathsf{NA}\}^p$ is a binary mask with NA denoting a missing value, and $\circ$ is the Hadamard (entrywise) product. Further, we consider a high-dimensional setting where n and p are growing with n possibly smaller than p.

We analyze the classical method of principal component regression (PCR) within this framework. PCR is a two-stage process: first, PCR "de-noises" the observed in-sample covariate matrix $\boldsymbol{Z} = [\boldsymbol{z}_i^\top] \in \mathbb{R}^{n \times p}$ via principal component analysis (PCA), i.e., PCR replaces $\boldsymbol{Z}$ by its low-rank approximation. Then, PCR regresses $\boldsymbol{y} = [y_i] \in \mathbb{R}^n$ on the low-rank approximation to produce the model estimate $\widehat{\boldsymbol{\beta}}$. The focus of this work is to answer the following questions about PCR:

Q1: "When $p > n$, is there a model parameter that PCR consistently identifies?"
Q2: "Given deterministic, corrupted, and partially observed out-of-sample covariates, can PCR recover the expected responses?"

1.1 Contributions

Model identification. Regarding Q1, we prove that PCR consistently identifies the projection of the model parameter onto the linear space generated by the underlying covariates. This corresponds to the unique minimum ℓ_2-norm model, which is arguably sufficient for valid statistical inference (Shao and Deng, 2012).

Out-of-sample prediction. For Q2, we leverage our results for Q1 to establish non-asymptotic out-of-sample prediction guarantees that improve upon the best known rates. Notably, these results are novel for the fixed design setting. In the course of our analysis, we introduce a natural linear algebraic condition between the in- and out-of-sample data that supplants distributional assumptions on the underlying covariates that are common in the literature. We construct a hypothesis test to check when this condition holds in practice. We also illustrate the importance of this condition through extensive simulations.

Applications to synthetic controls. Our responses to Q1–Q2 lead to novel results for the synthetic controls literature, a popular framework for policy evaluation (Abadie and Gardeazabal, 2003; Abadie et al., 2010). In particular, our results provide theoretical guarantees for several PCR based methods, namely Amjad et al. (2018, 2019). We apply our hypothesis test to two widely analyzed studies in the synthetic controls literature.

1.2 Organization

Section 2 details the PCR algorithm. Section 3 describes our problem setup and assumptions. Section 4 provides formal statistical guarantees on Q1–Q2. Section 5 reports on simulation studies. Section 6 presents a hypothesis test to check when a key assumption that enables PCR to generalize holds in practice. Section 7 contextualizes our findings within the synthetic controls framework. Section 8 discusses related works from the error-in-variables, PCR, and functional PCA/PCR literatures. Section 9 offers directions for future research. We relegate all mathematical proofs to the Appendix.

1.3 Notation

For a matrix $\boldsymbol{A} \in \mathbb{R}^{a \times b}$, we denote its operator (spectral), Frobenius, and max element-wise norms as $\|\boldsymbol{A}\|_2$, $\|\boldsymbol{A}\|_F$, and $\|\boldsymbol{A}\|_{\max}$, respectively. By rowspan($\boldsymbol{A}$), we denote the subspace of $\mathbb{R}^b$ spanned by the rows of $\boldsymbol{A}$. Let $\boldsymbol{A}^\dagger$ denote the pseudoinverse of $\boldsymbol{A}$. For a vector $\boldsymbol{v} \in \mathbb{R}^a$, let $\|\boldsymbol{v}\|_p$ denote its ℓ_p-norm for $p \in [1, \infty]$. We define the sub-gaussian (Orlicz) norm as $\|\boldsymbol{v}\|_{\psi_2}$. Let $\langle \cdot, \cdot \rangle$ and $\otimes$ denote the inner and outer products, respectively. For any two numbers $a, b \in \mathbb{R}$, we use $a \wedge b$ to denote $\min(a, b)$ and $a \vee b$ to denote $\max(a, b)$. Let $[a] = \{1, \ldots, a\}$ for any positive integer a.

Let f and g be two functions defined on the same space. We say that $f(n) = O(g(n))$ if and only if there exists a positive real number M and a real number n_0 such that for all $n \geq n_0, |f(n)| \leq M|g(n)|$. Analogously we say $f(n) = \Theta(g(n))$ if and only if there exists positive real numbers m, M such that for all $n \geq n_0,\ m|g(n)| \leq |f(n)| \leq M|g(n)|$; $f(n) = o(g(n))$ if for any $m > 0$, there exists n_0 such that for all $n \geq n_0, |f(n)| \leq m|g(n)|$; $f(n) = \omega(g(n))$ if for any $m > 0$, there exists n_0 such that for all $n \geq n_0, |f(n)| \geq m|g(n)|$. $\widetilde{O}(\cdot)$ is defined analogously to $O(\cdot)$, but ignores log dependencies.

2. Principal Component Regression

2.1 Observations

As described in Section 1, our in-sample (train) data consists of n *labeled* observations $\{(y_i, \boldsymbol{z}_i) : i \leq n\}$. By contrast, our out-of-sample (test) data consists of $m \geq 1$ *unlabeled* observations. That is, for $i > n$, we observe the covariates $\boldsymbol{z}_i$ but do not observe the associated response variables y_i. Let $\boldsymbol{Z} = [\boldsymbol{z}_i^\top : i \leq n] \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Z}' = [\boldsymbol{z}_i^\top : i > n] \in \mathbb{R}^{m \times p}$ denote the matrices of in- and out-of-sample covariates, respectively.

2.2 Description of Algorithm

We describe PCR, as introduced in Jolliffe (1982), with a variation to handle missing data.

I: Parameter estimation. Let $\widehat{\rho}$ denote the fraction of observed entries in $\boldsymbol{Z}$. Replace all missing values (NA) in the covariate matrices with zero. Let $\widetilde{\boldsymbol{Z}} = (1/\widehat{\rho})\boldsymbol{Z} = \sum_{i=1}^{n \wedge p} \widehat{s}_i \widehat{\boldsymbol{u}}_i \otimes \widehat{\boldsymbol{v}}_i$, where $\widehat{s}_i \in \mathbb{R}$ are the singular values (in decreasing order) and $\widehat{\boldsymbol{u}}_i \in \mathbb{R}^n, \widehat{\boldsymbol{v}}_i \in \mathbb{R}^p$ are the corresponding left and right singular vectors, respectively. For a hyperparameter $k \in [n \wedge p]$, let $\widetilde{\boldsymbol{Z}}^k = \sum_{i=1}^k \widehat{s}_i \widehat{\boldsymbol{u}}_i \otimes \widehat{\boldsymbol{v}}_i$ and define the estimated model parameter as

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{Z}}^k)^\dagger \boldsymbol{y} = \Big(\sum_{i=1}^k (1/\widehat{s}_i) \widehat{\boldsymbol{v}}_i \otimes \widehat{\boldsymbol{u}}_i \Big) \boldsymbol{y}. \tag{3}$$

II: Out-of-sample prediction. Let $\widehat{\rho}'$ denote the proportion of observed entries in $\boldsymbol{Z}'$. Let $\widetilde{\boldsymbol{Z}}' = (1/\widehat{\rho}')\boldsymbol{Z}' = \sum_{i=1}^{m \wedge p} \widehat{s}'_i \widehat{\boldsymbol{u}}'_i \otimes \widehat{\boldsymbol{v}}'_i$, where $\widehat{s}'_i \in \mathbb{R}$ are the singular values (in decreasing order) and $\widehat{\boldsymbol{u}}'_i \in \mathbb{R}^m, \widehat{\boldsymbol{v}}'_i \in \mathbb{R}^p$ are the left and right singular vectors, respectively. Given $\ell \in [m \wedge p]$, let $\widetilde{\boldsymbol{Z}}'^\ell = \sum_{i=1}^\ell \widehat{s}'_i \widehat{\boldsymbol{u}}'_i \otimes \widehat{\boldsymbol{v}}'_i$, and define the test response estimates as $\widehat{\boldsymbol{y}}' = \widetilde{\boldsymbol{Z}}'^\ell \widehat{\boldsymbol{\beta}}$.

If the expected responses are known to belong to a bounded interval, say $[-b, b]$ for some $b > 0$, then the entries of $\widehat{\boldsymbol{y}}'$ are truncated as follows: for every $i > n$,

$$\widehat{y}_i^{\text{trunc}} = \begin{cases} -b & \text{if } \widehat{y}_i < -b, \\ \widehat{y}_i & \text{if } -b \le \widehat{y}_i \le b, \\ b & \text{if } \widehat{y}_i > b. \end{cases} \tag{4}$$

2.3 Additional Useful Properties of PCR

We state a few useful properties of PCR that we use extensively. These are well-known results that are discussed in (Roman, 2008, Chapter 17) and (Strang, 2006, Chapter 6.3).

Property 2.1 *The PCR solution, $\widehat{\boldsymbol{\beta}}$, as given in* (3)*, is*

1. *the unique solution to the following program:*

$$\textit{minimize} \quad \|\boldsymbol{\beta}\|_2 \quad \textit{over} \quad \boldsymbol{\beta} \in \mathbb{R}^p \quad \textit{such that} \quad \boldsymbol{\beta} \in \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \ \|\boldsymbol{y} - \widetilde{\boldsymbol{Z}}^k \boldsymbol{b}\|_2^2.$$

2. *embedded within the* $\text{rowspan}(\widetilde{\boldsymbol{Z}}^k)$.

2.4 Applying PCR in Practice

2.4.1 Imputing Missing Covariate Values

As shown in Agarwal et al. (2021), PCR can equivalently be interpreted as first applying the matrix completion algorithm, hard singular value thresholding (HSVT), on $\widetilde{\boldsymbol{Z}}$ to obtain $\widetilde{\boldsymbol{Z}}^k$, and then performing OLS with this de-noised output matrix. Accordingly, this work utilizes the simple imputation method of replacing NA values with zero to enable HSVT. We justify this imputation approach as follows: by setting NA values to zero, it follows that $\mathbb{E}[Z_{ij}] = \rho X_{ij} + (1-\rho)0 = \rho X_{ij}$; recalling $\tilde{Z}_{ij} = (1/\widehat{\rho}) Z_{ij}$, we then obtain $\mathbb{E}[\tilde{Z}_{ij}] = X_{ij}$. Indeed, constructing $\widetilde{\boldsymbol{Z}}$ such that $\mathbb{E}[\widetilde{\boldsymbol{Z}}] = \boldsymbol{X}$ is a crucial step that enables the HSVT subroutine of PCR to produce a good estimate of $\boldsymbol{X}$ through $\widetilde{\boldsymbol{Z}}^k$.

Naturally, there are other matrix completion methods such as nearest neighbors or alternating least squares that do not first impute missing values. As long as the approach taken yields a sufficiently good estimator, cf. Lemma 3 of Appendix C, our main results on parameter estimation and generalization would naturally extend to these settings.

2.4.2 Choosing the Number of Principal Components

The ideal number of principal components k is rarely known a priori. As such, the problem of choosing k has become a well-studied problem in the low-rank matrix completion literature and there exists a suite of principled methods. These include visual inspections of the plotted singular values (Cattell, 1966), cross-validation (Wold, 1978; Owen and Perry, 2009), Bayesian methods (Hoff, 2007), and "universal" thresholding schemes that preserve singular values above a precomputed threshold (Gavish and Donoho, 2014; Chatterjee, 2015).

A common argument for these approaches is rooted in the underlying assumption that the smallest non-zero singular value of the "signal" $\boldsymbol{X}$ is well-separated from the largest

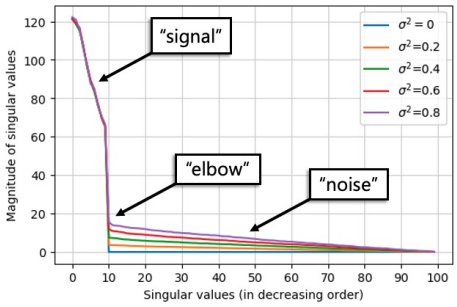

Figure 1: Spectrum of $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W} \in \mathbb{R}^{100\times 100}$. Here, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$, where entries of $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{100\times 10}$ are sampled independently from $\mathcal{N}(0,1)$; entries of $\boldsymbol{W}$ are sampled independently from $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 \in \{0, 0.2, \ldots, 0.8\}$. We see a steep drop-off in magnitude in the singular values across all noise levels—this marks the "elbow" point. Top singular values of $\boldsymbol{Z}$ correspond closely with those of $\boldsymbol{X}$ ($\sigma^2 = 0$). The remaining singular values are induced by $\boldsymbol{W}$. Thus, the "effective rank" of $\boldsymbol{Z}$ is the rank of $\boldsymbol{X}$.

singular value of the "noise" $\boldsymbol{W}$. Under reasonable "signal-to-noise" (snr) scenarios, Weyl's inequality implies that a sharp threshold or gap should exist between the top $r = \text{rank}(\boldsymbol{X})$ singular values and remaining singular values of the observed data $\widetilde{\boldsymbol{Z}}$. This gives rise to a natural "elbow" point, shown in Figure 1, and suggests choosing a threshold within this gap. As such, a researcher can simply plot the singular values of $\widetilde{\boldsymbol{Z}}$ and look for the elbow structure to decide if PCR is suitable for the application at hand. We formalize a notion of snr in (5) and establish our results in the following section with respect to this quantity.

3. Problem Setup

This section formalizes our problem setup. Let $\boldsymbol{X} = [\boldsymbol{x}_i^\top : i \le n] \in \mathbb{R}^{n\times p}$ and $\boldsymbol{X}' = [\boldsymbol{x}_i^\top : i > n] \in \mathbb{R}^{m\times p}$ represent the underlying in- and out-of-sample covariates, respectively.

3.1 Assumptions

Collectively, we assume (1) and (2) are satisfied. We make the additional assumptions.

Assumption 3.1 (Response noise) *Let $\{\boldsymbol{\varepsilon}_i : i \le n\}$ be a sequence of independent mean zero subgaussian random variables with $\|\boldsymbol{\varepsilon}_i\|_{\psi_2} \le \sigma$.*

Assumption 3.1 is a standard assumption in the regression literature that posits the idiosyncratic response noise to be independent draws from a subgaussian distribution.

Assumption 3.2 (Covariate noise and missing values) *Let $\{\boldsymbol{w}_i : i \le n + m\}$ be a sequence of independent mean zero subgaussian random vectors with $\|\boldsymbol{w}_i\|_{\psi_2} \le K$ and $\|\mathbb{E}[\boldsymbol{w}_i \otimes \boldsymbol{w}_i]\|_2 \le \gamma^2$. Let $\boldsymbol{\pi}_i \in \{1, \mathsf{NA}\}^p$, where NA denotes a missing value, be a vector of independent Bernoulli variables with parameter $\rho \in (0, 1]$. Further, let $\boldsymbol{\varepsilon}_i$, $\boldsymbol{w}_i$, $\boldsymbol{\pi}_i$ be mutually independent.*

Consistent with standard assumptions in the error-in-variables (EiV) regression literature, Assumption 3.2 posits the idiosyncratic EiV vector-valued noise $\boldsymbol{w}_i$ to be subgaussian and

independent across measurements; note, however, that the noise is allowed to be dependent within a measurement, i.e., the coordinates of $\boldsymbol{w}_i$ can be correlated. Finally, we require missing entries in the observed covariate vector to be missing completely at random (MCAR). In Section 4.3.1, we discuss ways to allow for more heterogeneous missingness patterns.

Assumption 3.3 (Bounded covariates) *Let* $\|\boldsymbol{X}\|_{\max} \leq 1$ *and* $\|\boldsymbol{X}'\|_{\max} \leq 1$.

Assumption 3.3 bounds the magnitude of the underlying noiseless covariates, not the *observed* noisy covariates. This assumption is made to simplify our analysis and it can be generalized to hold for any C that is an absolute constant. Our theoretical results will correspondingly only change by an absolute constant as well.

4. Main Results

This section addresses Q1–Q2. For ease of notation, let $C, c > 0$ be absolute constants whose values may change from line to line or even within a line. Let $\boldsymbol{H} = \boldsymbol{X}^\dagger \boldsymbol{X} \in \mathbb{R}^{p \times p}$ and $\boldsymbol{H}_\perp = \boldsymbol{I} - \boldsymbol{H}$ denote the projection matrices onto the rowspace and nullspace of $\boldsymbol{X}$, respectively. Let $\boldsymbol{H}', \boldsymbol{H}'_\perp \in \mathbb{R}^{p \times p}$ be defined analogously with respect to $\boldsymbol{X}'$. We define $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{H}\boldsymbol{\beta}^*$ as the projection of $\boldsymbol{\beta}^*$ onto the linear space spanned by the rows of $\boldsymbol{X}$.

4.1 Model Identification

Q1: "When $p > n$, is there a model parameter that PCR consistently identifies?"

The model parameter $\boldsymbol{\beta}^*$ is not identifiable in the high-dimensional regime as infinitely many solutions satisfy (1). Among all feasible parameters, we show that PCR recovers $\tilde{\boldsymbol{\beta}}^*$, the unique parameter with minimum ℓ_2-norm that is entirely embedded in the rowspace of $\boldsymbol{X}$, provided the number of principal components k is aptly chosen.

From Property 2.1, recall that PCR enforces $\widehat{\boldsymbol{\beta}} \in \text{rowspan}(\widetilde{\boldsymbol{Z}}^k)$. Hence, if $k = r = \text{rank}(\boldsymbol{X})$ and the rowspace of $\widetilde{\boldsymbol{Z}}^r$ is "close" to the rowspace of $\boldsymbol{X}$, then $\widehat{\boldsymbol{\beta}} \approx \tilde{\boldsymbol{\beta}}^*$. The "noise" in $\boldsymbol{Z}$ arises from the missingness pattern induced by $\boldsymbol{\pi}$ and the measurement error $\boldsymbol{W}$; meanwhile, the "signal" in $\boldsymbol{Z}$ arises from $\boldsymbol{X}$, where its strength is captured by the magnitude of its singular values. Accordingly, we define the snr as

$$\mathsf{snr} := \frac{\rho s_r}{\sqrt{n} + \sqrt{p}}. \tag{5}$$

Here, s_r is the smallest nonzero singular value of $\boldsymbol{X}$, ρ determines the fraction of observed entries, and $\sqrt{n} + \sqrt{p}$ is induced by the perturbation in the singular values from $\boldsymbol{W}$. As one would expect, the signal strength snr scales linearly with ρ. From standard concentration results for sub-gaussian matrices, it follows that $\|\boldsymbol{W}\|_2 = \widetilde{O}(\sqrt{n} + \sqrt{p})$ (see Lemma 9). With this notation, we state the main result on model identification.

Theorem 4.1 *Let Assumptions 3.1–3.3 hold. Consider (i) PCR with* $k = r = \text{rank}(\boldsymbol{X})$, *(ii)* $\rho \geq c(np)^{-1}\log^2(np)$, *and (iii)* $\mathsf{snr} \geq C(K+1)(\gamma+1)$. *Then w.p. at least* $1 - O((np)^{-10})$,

$$\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \leq C_{\text{noise}} \log(np) \cdot \left\{ \frac{\|\tilde{\boldsymbol{\beta}}^*\|_2^2}{\mathsf{snr}^2} + \frac{\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1}{(n \vee p)\mathsf{snr}^2} + \frac{r(1 \vee \|\tilde{\boldsymbol{\beta}}^*\|_1^2)}{(n \vee p)\mathsf{snr}^2} + \frac{\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^4} \right\}, \tag{6}$$

where $C_{\text{noise}} = C(K+1)^4(\gamma+1)^2(\sigma^2+1)$. *Further, if* $\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle \in [-d, d]$ *for all* $i \leq n$, *then*

$$\|\tilde{\boldsymbol{\beta}}^*\|_2 \leq s_r^{-1} \cdot d\sqrt{n} \quad \textit{and} \quad \|\tilde{\boldsymbol{\beta}}^*\|_1 \leq s_r^{-1} \cdot d\sqrt{np}. \tag{7}$$

Interpretation. Let us discuss why the ℓ_1-norm of $\tilde{\boldsymbol{\beta}}^*$ appears in the bound. Our analysis of the parameter estimation error involves an EiV error term of the form $\|(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k)\tilde{\boldsymbol{\beta}}^*\|_2$, which can be bounded as follows:

$$\|(\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k)\tilde{\boldsymbol{\beta}}^*\|_2 \leq \|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}\|\tilde{\boldsymbol{\beta}}^*\|_1,$$

where $\|\boldsymbol{A}\|_{2,\infty}^2 = \max_{j\in[p]} \sum_{i=1}^n A_{ij}^2$ for any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$; see (S8) in Appendix C for details. From (7), it is clear that $\|\tilde{\boldsymbol{\beta}}^*\|_1$ is controlled if s_r is sufficiently large. Indeed, Assumption 4.1 below is one such natural condition on s_r.

To gain a better view on Theorem 4.1 regarding consistency, let us suppress dependencies on (K, γ, σ) for the following discussion. Theorem 4.1 implies that a sufficient condition for consistency is given by

$$\frac{\mathsf{snr}^2}{\log(np) \cdot \max\{\|\tilde{\boldsymbol{\beta}}^*\|_2^2,\ n^{1/2}(n\vee p)^{-1}\|\tilde{\boldsymbol{\beta}}^*\|_1,\ r(n\vee p)^{-1}(1\vee\|\tilde{\boldsymbol{\beta}}^*\|_1^2),\ \|\tilde{\boldsymbol{\beta}}^*\|_1\}} \to \infty.$$

That is, PCR recovers $\tilde{\boldsymbol{\beta}}^*$ provided snr grows sufficiently fast. Finally, (7) implies that (6) can be purely expressed through the smallest nonzero singular value of $\boldsymbol{X}$.

We now describe a natural setting for which we can provide an explicit bound on the snr. Towards this, we introduce the following assumption and discuss its meaning in Section 8.3.

Assumption 4.1 (Balanced spectra: in-sample covariates) *The* r *nonzero singular values* s_i *of* $\boldsymbol{X}$ *satisfy* $s_i = \Theta(\sqrt{np/r})$.

Corollary 4.1 *Let the setup of Theorem 4.1 and Assumption 4.1 hold. If* $\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle \in [-d, d]$ *for all* $i \leq n$, *then w.p. at least* $1 - O((np)^{-10})$,

$$\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \leq C_{\text{noise}} \log(np) \cdot \left\{ \frac{dr^{3/2}}{\rho^2\sqrt{np}} + \frac{d^2r^3}{\rho^4(n\wedge p)^2} \right\}.$$

Proof By Assumption 4.1, we have $s_r = \Theta(\sqrt{np/r})$. This yields

$$\mathsf{snr} = \frac{\rho s_r}{\sqrt{n}+\sqrt{p}} \ \geq \frac{c\rho\sqrt{np}}{\sqrt{r(n+p)}} \ \geq c\rho\sqrt{(n\wedge p)/r},$$

i.e., $\mathsf{snr} = \Omega(\rho\sqrt{(n\wedge p)/r})$. Further, we have from (7) that

$$\|\tilde{\boldsymbol{\beta}}^*\|_2 \leq d\sqrt{r/p}, \quad \|\tilde{\boldsymbol{\beta}}^*\|_1 \leq d\sqrt{r}. \tag{8}$$

Inserting (8) into (6) and simplifying completes the proof. ■

Ignoring dependencies on (ρ, r, d), Corollary 4.1 implies that the model identification error scales as $\min\{1/\sqrt{np}, 1/(n\wedge p)^2\}$. Hence, the error vanishes as $\min\{n, p\} \to \infty$. We note that the dependency on $\min\{n, p\}$ arises from the EiV problem, i.e., the error incurred from estimating the subspaces spanned by the left and right singular vectors of $\boldsymbol{X}$ through the corresponding singular vectors of $\widetilde{\boldsymbol{Z}}$; en route to our end result, we show that $\widetilde{\boldsymbol{Z}}^k$ is a good estimate of $\boldsymbol{X}$ provided both n and p grow (see Lemmas 2 and 3 in Appendix C for details).

4.2 Out-of-sample Prediction

Q2: "Given deterministic, corrupted, and partially observed out-of-sample covariates, can PCR recover the expected responses?"

Towards answering Q2, we define PCR's out-of-sample (test) prediction errors with respect to $\widehat{\boldsymbol{y}}$ and $\widehat{\boldsymbol{y}}^{\text{trunc}}$ as

$$\text{MSE}_{\text{test}} \coloneqq \frac{1}{m}\sum_{i=1}^{m}(\widehat{y}_{n+i} - \langle \boldsymbol{x}_{n+i}, \boldsymbol{\beta}^* \rangle)^2 \tag{9}$$
$$\text{MSE}^{\text{trunc}}_{\text{test}} \coloneqq \frac{1}{m}\sum_{i=1}^{m}(\widehat{y}^{\text{trunc}}_{n+i} - \langle \boldsymbol{x}_{n+i}, \boldsymbol{\beta}^* \rangle)^2,$$

respectively, where we recall that $\widehat{y}^{\text{trunc}}$, as given by (4), implicitly depends on the known bound, b. Let $s_\ell, s'_\ell \in \mathbb{R}$ be the ℓ-th largest singular values of $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. Recall from Section 2 that $\widehat{s}_\ell, \widehat{s}'_\ell$ are defined analogously for $\widetilde{\boldsymbol{Z}}$ and $\widetilde{\boldsymbol{Z}}'$, respectively. Analogous to (5), we define a signal-to-noise ratio for the out-of-sample covariates as

$$\mathsf{snr}_{\text{test}} \coloneqq \frac{\rho s'_{r'}}{\sqrt{m} + \sqrt{p}}. \tag{10}$$

Next, we bound MSE_{test} in probability and $\text{MSE}^{\text{trunc}}_{\text{test}}$ in expectation with respect to snr and $\mathsf{snr}_{\text{test}}$. For ease of notation, we define $n_{\min} = n \wedge m$ and $n_{\max} = n \vee m$.

Theorem 4.2 *Let the setup of Theorem 4.1 hold with $\rho \geq c(mp)^{-1}\log^2(mp)$. Consider (i) PCR with $\ell = r' = \text{rank}(\boldsymbol{X}')$ and (ii) $\|\tilde{\boldsymbol{\beta}}^*\|_1 = \Omega(1)$. Then w.p. at least $1 - O((n_{\min}p)^{-10})$,*

$$\text{MSE}_{\text{test}} \leq \Delta_1 + \Delta_2, \tag{11}$$

where

$$\Delta_1 = C \cdot p \cdot \delta_\beta \cdot \|\boldsymbol{H}_\perp \boldsymbol{H}'\|_2^2,$$
$$\Delta_2 = C'_{\text{noise}} \log(n_{\max}p) \cdot \left\{ \frac{\sqrt{n}}{\mathsf{snr}^2}\|\tilde{\boldsymbol{\beta}}^*\|_1 + \Delta_3 \right\},$$
$$\Delta_3 = \left(\frac{r(1 \vee p/m)}{\rho^2 \mathsf{snr}^2} + \frac{r'}{\mathsf{snr}^2_{\text{test}} \wedge m} + \frac{n \vee p}{\mathsf{snr}^4} \right) \|\tilde{\boldsymbol{\beta}}^*\|_1^2;$$

here, δ_β is given by the right-hand side of (6) and $C'_{\text{noise}} = C(K+1)^6(\gamma+1)^4(\sigma^2+1)$. Further, if $\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle \in [-b, b]$ for all $i > n$, then

$$\mathbb{E}[\text{MSE}^{\text{trunc}}_{\text{test}}] \leq \Delta_1 + \Delta_4 + \Delta_5, \tag{12}$$

where

$$\Delta_4 = C'_{\text{noise}} \log(n_{\max}p) \cdot \left\{ \frac{\sqrt{n}}{\mathsf{snr}^2} \left(\frac{1}{\mathsf{snr}^2} + \frac{1 \vee p/m}{\rho^2(n \vee p)} \right) \|\tilde{\boldsymbol{\beta}}^*\|_1 + \Delta_3 \right\},$$
$$\Delta_5 = \frac{Cb^2}{(n_{\min}p)^{10}}.$$

Interpretation. Let us briefly dissect Theorem 4.2. Firstly, condition (ii) is not necessary but made to simplify the resulting bound. On a more interesting note, it is well known that generalization error bounds rely on some notion of "closeness" between the in- and out-of-sample covariates. A canonical assumption within the statistical learning theory literature considers the two sets of covariates to be drawn from the same underlying distribution à la i.i.d. samples. As seen in (11) and (12), we consider a complementary notion of covariate closeness that is captured by the term $\|\boldsymbol{H}_{\perp}\boldsymbol{H}'\|_2$ in Δ_1. In words, it measures the size of the linear subspace spanned by the out-of-sample covariates that is not contained within the linear subspace spanned by the in-sample covariates. Effectively, this term quantifies the ℓ_2-distance, or ℓ_2-similarity, between the in- and out-of-sample covariates. If each out-of-sample covariate is some linear combination of the in-sample covariates, then this error term vanishes and the out-of-sample prediction error decreases. We formalize this concept in Assumption 4.2 below.

Assumption 4.2 (Subspace inclusion) *Let* $\text{rowspan}(\boldsymbol{X}') \subseteq \text{rowspan}(\boldsymbol{X})$.

To aid our intuition of Assumption 4.2, consider (1) in the classical regime where $n > p$. The canonical assumption within this paradigm considers $\boldsymbol{X}$ to have full column rank, i.e., $\text{rank}(\boldsymbol{X}) = p$. Accordingly, the in-sample covariates span $\mathbb{R}^p$ so the subspace spanned by the out-of-sample covariates necessarily lies within that spanned by the in-sample covariates, yielding $\|\boldsymbol{H}_{\perp}\boldsymbol{H}'\|_2 = 0$. In this view, Assumption 4.2 generalizes the full column rank assumption in the classical regime to the collinear setting in the high-dimensional regime.

Corollary 4.2 *Let the setup of Theorem 4.2 and Assumption 4.2 hold. Then,* $\Delta_1 = 0$.

Proof Under Assumption 4.2, we have $\|\boldsymbol{H}'\boldsymbol{H}_{\perp}\|_2^2 = 0$. ■

For interpretability, we suppress dependencies on (K, γ, σ), and assume $p = \Theta(m)$ with $m \to \infty$. One can then verify that Corollary 4.2 implies that sufficient conditions for PCR's expected test prediction error to vanish are

$$\frac{\mathsf{snr}^2}{\log(n_{\max}p) \cdot \max\{n^{1/4}\|\tilde{\boldsymbol{\beta}}^*\|_1^{1/2},\ (n \vee p)^{1/2}\|\tilde{\boldsymbol{\beta}}^*\|_1\}} \to \infty,$$
$$\frac{\rho^2\ \mathsf{snr}^2}{\log(n_{\max}p) \cdot \max\{\ n^{1/2}(1 \vee p/m)(n \vee p)^{-1}\|\tilde{\boldsymbol{\beta}}^*\|_1,\ r(1 \vee p/m)\|\tilde{\boldsymbol{\beta}}^*\|_1^2\}} \to \infty,$$
$$\frac{\mathsf{snr}_{\text{test}}^2}{\log(n_{\max}p) \cdot r'\|\tilde{\boldsymbol{\beta}}^*\|_1^2} \to \infty.$$

As with Theorem 4.1, we specialize Theorem 4.2 in Corollary 4.3 to the setting where $\mathsf{snr} = \Omega(\rho\sqrt{(n \wedge p)/r})$ and $\mathsf{snr}_{\text{test}} = \Omega(\rho\sqrt{(m \wedge p)/r'})$. A sufficient condition for the lower bound on $\mathsf{snr}_{\text{test}}$ is provided in Assumption 4.3.

Assumption 4.3 (Balanced spectra: out-of-sample covariates) *The* r' *nonzero singular values* s_i' *of* $\boldsymbol{X}'$ *satisfy* $s_i' = \Theta(\sqrt{mp/r'})$.

Corollary 4.3 *Let the setups of Corollaries 4.1–4.2 and Assumption 4.3 hold. Then w.p. at least* $1 - O((n_{\min}p)^{-10})$,

$$\mathrm{MSE}_{\mathrm{test}} \leq C'_{\mathrm{noise}} \log(n_{\max}p) \cdot \left\{ \frac{dr^{3/2}\sqrt{n}}{\rho^2(n \wedge p)} + \Delta \right\},$$

where

$$\Delta = \frac{d^2r^3(1 \vee p/m)}{\rho^4(n \wedge p)} + \frac{d^2r^2}{\rho^2(m \wedge p)} + \frac{d^2r(n \vee p)}{\rho^4(n \wedge p)^2}.$$

Further, if $\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle \in [-b, b]$ *for all* $i > n$, *then*

$$\mathbb{E}[\mathrm{MSE}^{\mathrm{trunc}}_{\mathrm{test}}] \leq C'_{\mathrm{noise}} \log(n_{\max}p) \cdot \left\{ \frac{dr^{5/2}\sqrt{n}}{\rho^4(n \wedge p)(n \wedge m \wedge p)} + \Delta \right\} + \frac{Cb^2}{(n_{\min}p)^{10}}.$$

Proof Using identical arguments to those made in the proof of Corollary 4.1 and noting $r' \leq r$, it follows that Assumption 4.3 gives $\mathsf{snr}_{\mathrm{test}} \geq c\rho\sqrt{(m \wedge p)/r}$. Plugging the bounds on snr, $\mathsf{snr}_{\mathrm{test}}$, and (8) into Theorem 4.2 completes the proof. ■

For the following discussion, we suppress dependencies on (K, γ, σ, r) and log factors, assume $\rho = \Theta(1)$, and only consider the scaling with respect to (n, m, p). Corollary 4.3 implies that if $p = o(nn_{\min})$ and $n = o(p^2)$,[1] then the out-of-sample prediction error vanishes to zero both in expectation and w.h.p., as $n, m, p \to \infty$. If we make the additional assumption that $n = \Theta(p)$ and $p = \Theta(m)$, then the error scales as $\widetilde{O}(1/n)$ in expectation. This improves upon the best known rate of $\widetilde{O}(1/\sqrt{n})$, established in Agarwal et al. (2021); notably, these works do not provide a high probability bound. Additionally, Agarwal et al. (2021) require i.i.d. covariates to leverage standard Rademacher tools for their out-of-sample analyses. In contrast, we consider fixed design points, thus our generalization error bounds do not rely on distributional assumptions regarding $\boldsymbol{X}$ and $\boldsymbol{X}'$. Finding the optimal relative scalings of (n, m, p) to achieve consistency remains future work.

4.3 Discussion

4.3.1 Heterogeneous Missingness Patterns

Assumption 3.2 considers MCAR patterns in the observed covariate matrix $\boldsymbol{Z}$. This is motivated by the HSVT subroutine of PCR, as discussed in Section 2.4.1. If the missingness pattern is instead heterogeneous, other matrix completion methods designed for such settings can be utilized to more accurately recover the underlying covariates. Matrix completion with heterogeneous missingness patterns is an active area of research and there has been a recent emergence of exciting results, including Schnabel et al. (2016); Ma and Chen (2019); Sportisse et al. (2020) and Bhattacharya and Chatterjee (2022) to name a few.

At a high-level, these algorithms follow a two-step approach: (i) construct estimates $\widehat{\rho}_{ij}$ of ρ_{ij}; (ii) use $\widehat{\rho}_{ij}$ and $\boldsymbol{Z}$ to estimate X_{ij}. With regards to step (i), let $\boldsymbol{\Pi} \in \{0, 1\}^{n \times p}$

1. Practically speaking, this condition is not binding. If $n = \Omega(p^2)$, then we can sample a subset of the training data to satisfy the condition. Hence, this condition is likely an artifact of our analysis.

denote the binary mask matrix with $\mathbb{E}[\pi_{ij}] = \rho_{ij}$. The common assumption driving these approaches is that $\mathbb{E}[\boldsymbol{\Pi}]$ is a low-rank matrix; note that the MCAR paradigm with $\mathbb{E}[\pi_{ij}] = \rho$ yields $\text{rank}(\mathbb{E}[\boldsymbol{\Pi}]) = 1$. As such, matrix completion algorithms can be first applied to $\boldsymbol{\Pi}$ to obtain the estimates $\widehat{\rho}_{ij}$. Then, $\boldsymbol{X}$ can be estimated using $\widehat{\rho}_{ij}$ and $\boldsymbol{Z}$. Within the context of this work, if the matrix completion algorithm can faithfully recover the underlying covariates, cf. Lemma 3 of Appendix C, then our main results in Section 4 would naturally extend. A formal analysis of this more general estimator is left as interesting future work.

For the specific setting where there is a different probability of missingness $\{\rho_j\}_{j\in[p]}$ for each of the p covariates, we propose a straightforward extension of PCR. Let $\widehat{\rho}_j$ denote the fraction of observed entries in the j-th column of $\boldsymbol{Z}$. Let $\widehat{\boldsymbol{P}} \in \mathbb{R}^{p\times p}$ be a diagonal matrix with the j-th diagonal element given by $\widehat{\rho}_j$. After setting the NA values of $\boldsymbol{Z}$ to zero, we now redefine $\widetilde{\boldsymbol{Z}}$ as $\widetilde{\boldsymbol{Z}} = \boldsymbol{Z}\widehat{\boldsymbol{P}}$. In words, rather than uniformly re-weighting the $\boldsymbol{Z}$ by $1/\widehat{\rho}$, we re-weigh the j-th column of $\boldsymbol{Z}$ by $1/\widehat{\rho}_j$. As a result, our theoretical results will go through in an analogous manner with the scaling now depending on $\rho_{\min} = \min_{j\in[p]} \rho_j$. It remains an interesting future direction to further refine this strategy such that the theory depends on the "average" missingness probability.

Implicitly, the approach proposed above as well as the standard toolkit of matrix completion methods hinge on two limiting assumptions (Ma and Chen, 2019): (i) *positivity*—each entry is observed with positive probability, i.e., $\min_{i,j} \rho_{ij} > 0$, and (ii) *independence*—each entry is observed independently of other entries. Violating these assumptions creates significant challenges in estimating the underlying probabilities and can preclude common methods such as those based on inverse propensity weighting. The challenge of estimating the observation probabilities is further exacerbated when the missingness pattern is correlated with the underlying matrix entries; this is commonly referred as the missing not at random paradigm or confounding (Rubin, 1976; Little and Rubin, 2019). A formal treatment on this subject is left as important future work.

4.3.2 PCR Theory with Misspecified Number of Principal Components

The results of this section rely on an oracle version of PCR that has access to the true ranks of $\boldsymbol{X}$ and $\boldsymbol{X}'$. We leave a formal treatment of PCR when the number of principal components is misspecified as an important future line of inquiry. With that said, we remark that the universal data-driven approach of Gavish and Donoho (2014), as mentioned in Section 2.4.2, often performs remarkably well in practice. We apply this approach in our simulation studies on PCR's generalization performance in Sections 5.2–5.4. We further present a simulation in Appendix B that studies PCR's predictive accuracy when k is misspecified; there, we find that PCR can still predict well when $k > r$ but suffers significantly when $k < r$, which suggests that practitioners should err on the side of including more principal components than less.

4.3.3 Towards a Lower Bound on Model Identification

To the best of our knowledge, Theorem 4.1 provides the first upper bound on PCR's model parameter estimation error in the high-dimensional EiV setting with fixed design. In Lemma 24 of Appendix G, we take the first step towards establishing a complementary lower bound to better understand the limitations of PCR in such a setting.

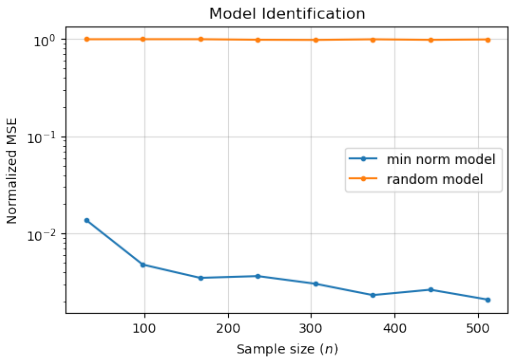

Figure 2: Plots of ℓ_2-norm errors (log scale) for two cases: (i) $\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2$ and (ii) $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. The error decays for case (i) but remains stagnant for case (ii).

4.3.4 Viewing Generalization through Assumption 4.2

As discussed, our out-of-sample guarantees do *not* rely on any distributional assumptions between the in- and out-of-sample covariates. Rather, our results rely on a purely linear algebraic condition given by Assumption 4.2. In this view, Assumption 4.2 offers a complementary, distribution-free perspective on generalization and has possible implications to learning under covariate shifts. We examine the role of Assumption 4.2 in our simulations in Section 5. As a preview, our results provide empirical evidence that PCR can generalize even when the in- and out-of-sample covariates obey different distributions provided Assumption 4.2 holds. In light of these findings, we furnish a data-driven diagnostic in Section 6 to check when Assumption 4.2 may hold in practice.

5. Illustrative Simulations

In this section, we present illustrative simulations to support our theoretical results. We provide details of the simulations in Appendix A.

5.1 PCR Identifies the Minimum ℓ_2-norm Model Parameter

To see how Theorem 4.1 plays out in practice, we design a simulation on model identification.

Setup. We consider $p = 512$ and $r = 15$. We generate $\boldsymbol{\beta}^*$ and set it to have unit norm. For each $n \in \{30, 98, 167, \ldots, p\}$, we generate the $\boldsymbol{X}$ and define the minimum ℓ_2-norm solution as $\tilde{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger \boldsymbol{X} \boldsymbol{\beta}^*$. We conduct 1000 simulation repeats per sample size n. For each repeat, we sample $(\boldsymbol{\varepsilon}, \boldsymbol{W})$ to construct $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$.

Results. For each simulation repeat, we apply PCR on $(\boldsymbol{y}, \boldsymbol{Z})$ to learn a *single* $\widehat{\boldsymbol{\beta}}$ with $k = r$ chosen correctly. Figure 2 visualizes the root-MSE (RMSE) of $\widehat{\boldsymbol{\beta}}$ with respect to $\tilde{\boldsymbol{\beta}}^*$ and $\boldsymbol{\beta}^*$. As predicted by Theorem 4.1, the RMSE with respect to $\tilde{\boldsymbol{\beta}}^*$ decays to zero as the sample size increases. In contrast, the RMSE with respect to $\boldsymbol{\beta}^*$ stays roughly constant across different sample sizes. This reaffirms that PCR identifies the minimum ℓ_2-norm solution amongst all feasible solutions.

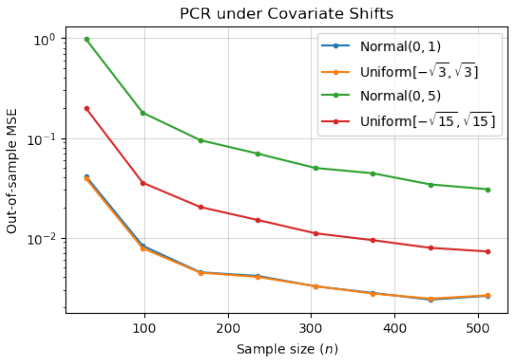

Figure 3: Plot of PCR's MSE (log scale) under various covariate shifts with Assumption 4.2 satisfied in each case. The MSE decays as the sample size increases for each covariate shift.

5.2 PCR is Robust to Covariate Shifts

We study the PCR's generalization properties, as predicted by Theorem 4.2, in the presence of covariate shifts, i.e., the in- and out-of-sample covariates follow different distributions.

Setup. We choose the same relative scalings of (n, p, r) and generate $\boldsymbol{\beta}^*$ as in Section 5.1. Let $m = n$. For each n, we generate $\boldsymbol{X}$ as per distribution $\mathcal{D}_1$. We then generate four different out-of-sample covariates as follows: (i) $\boldsymbol{X}_1' \sim \mathcal{D}_1$, (ii) $\boldsymbol{X}_2' \sim \mathcal{D}_2$, (iii) $\boldsymbol{X}_3' \sim \mathcal{D}_3$, and (iv) $\boldsymbol{X}_4' \sim \mathcal{D}_4$, where $\mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4$ are distinct distributions from one another and from $\mathcal{D}_1$. Critically, Assumption 4.2 is satisfied between $\boldsymbol{X}$ and $\boldsymbol{X}_i'$ for every $i \in \{1, \ldots, 4\}$. We define $\boldsymbol{\theta}_i' = \boldsymbol{X}_i'\boldsymbol{\beta}^*$. We conduct 1000 simulation repeats. For each repeat, we sample $(\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{W}')$ to construct $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, and $\boldsymbol{Z}_i' = \boldsymbol{X}_i' + \boldsymbol{W}'$.

Results. For each simulation repeat, we apply PCR on $(\boldsymbol{y}, \boldsymbol{Z})$ to learn a single $\widehat{\boldsymbol{\beta}}$ by choosing k via the universal data-driven approach of Gavish and Donoho (2014). For each i, we construct $\widehat{\boldsymbol{y}}_i'$ from the de-noised version of $\boldsymbol{Z}_i'$ and $\widehat{\boldsymbol{\beta}}$. Figure 3 displays the MSE of $\widehat{\boldsymbol{y}}_i'$ with respect to $\boldsymbol{\theta}_i'$. As predicted by Corollary 4.3, the out-of-sample prediction error decays as the sample size increases for each covariate shift. Hence, our results provide further evidence that PCR is robust to corrupted out-of-sample covariates and, perhaps more importantly, covariate shifts provided Assumption 4.2 holds.

5.3 PCR Generalizes under Assumption 4.2

This simulation further examines the role of Assumption 4.2. Specifically, we compare PCR's generalization error under two settings: (i) there is covariate shift but Assumption 4.2 holds; (ii) there is distributional invariance (i.e., the in- and out-of-sample covariates obey the same distribution) but Assumption 4.2 is violated.

Setup. We choose the same relative scalings of (n, m, p, r) and generate $\boldsymbol{\beta}^*$ as in Section 5.2. For each n, we generate $\boldsymbol{X} \sim \mathcal{D}_1$. We then generate two out-of-sample covariates: (i) $\boldsymbol{X}_1' \sim \mathcal{D}_1$ that violates Assumption 4.2; (ii) $\boldsymbol{X}_2' \sim \mathcal{D}_2$ with $\mathcal{D}_2 \neq \mathcal{D}_1$ that obeys Assumption 4.2. Next, we define $\boldsymbol{\theta}_i' = \boldsymbol{X}_i'\boldsymbol{\beta}^*$ for $i \in \{1, 2\}$. We conduct 1000 simulation repeats. For each repeat, we sample $(\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{W}')$ to construct $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, and $\boldsymbol{Z}_i' = \boldsymbol{X}_i' + \boldsymbol{W}'$.

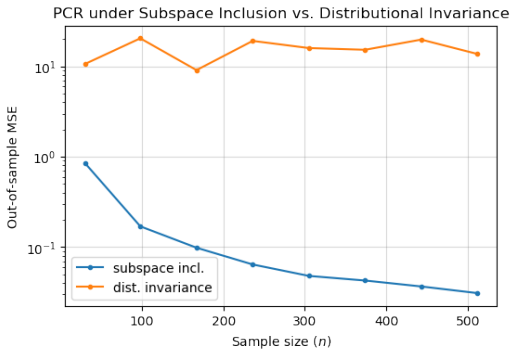

Figure 4: Plots of PCR's MSE (log scale) under two cases: (i) Assumption 4.2 holds but distributional invariance is violated (blue); (ii) Assumption 4.2 is violated but distributional invariance holds (orange). Case (i) achieves a vanishing MSE while case (ii) suffers from non-vanishing MSE.

Results. For each simulation repeat, we apply PCR on $(\boldsymbol{y}, \boldsymbol{Z})$ to learn a single $\widehat{\boldsymbol{\beta}}$ by choosing k via the universal data-driven approach of Gavish and Donoho (2014). For each i, we construct $\widehat{\boldsymbol{y}}_i'$ from the de-noised version of $\boldsymbol{Z}_i'$ and $\widehat{\boldsymbol{\beta}}$. Figure 4 displays the MSE of $\widehat{\boldsymbol{y}}_i'$ with respect to $\boldsymbol{\theta}_i'$. When Assumption 4.2 holds, the MSE decays as the sample size increases; by contrast, when Assumption 4.2 fails, the MSE is stagnant across varying sample sizes. Our findings reinforce the importance of Assumption 4.2 for PCR's ability to generalize.

5.4 PCR Generalizes with MCAR Entries

This simulation investigates PCR's out-of-sample performance under varying intensities of MCAR patterns in the observed covariate matrices.

Setup. We choose the same relative scalings of (n, m, p, r) and generate $\boldsymbol{\beta}^*$ as in Section 5.2. For each n, we generate $\boldsymbol{X}, \boldsymbol{X}' \sim \mathcal{D}_1$ with Assumption 4.2 satisfied. Next, we define $\boldsymbol{\theta}' = \boldsymbol{X}'\boldsymbol{\beta}^*$. We consider varying intensities of MCAR entries with $\rho \in \{0.4, 0.6, 0.8, 0.99\}$. We conduct 1000 simulation repeats for each (ρ, n) pair. For each repeat, we sample $(\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{W}', \boldsymbol{\Pi}, \boldsymbol{\Pi}')$ to construct $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{Z} = (\boldsymbol{X} + \boldsymbol{W}) \circ \boldsymbol{\Pi}$, and $\boldsymbol{Z}_i' = (\boldsymbol{X}_i' + \boldsymbol{W}') \circ \boldsymbol{\Pi}'$. Note that there are ρ entries in $\boldsymbol{\Pi}, \boldsymbol{\Pi}'$ that are randomly assigned the value 1, and each iteration considers a different permutation of revealed entries.

Results. For each simulation repeat, we apply PCR on $(\boldsymbol{y}, \boldsymbol{Z})$ to learn $\widehat{\boldsymbol{\beta}}$ by choosing k via the universal data-driven approach of Gavish and Donoho (2014). We construct $\widehat{\boldsymbol{y}}'$ from the de-noised version of $\boldsymbol{Z}'$ and $\widehat{\boldsymbol{\beta}}$. Figure 5 displays the MSE of $\widehat{\boldsymbol{y}}'$ with respect to $\boldsymbol{\theta}'$. Across varying intensities of ρ, the MSE decays as the sample size increases, which suggests that PCR can generalize when entries in the observed covariate matrices are MCAR.

6. A Hypothesis Test for Assumption 4.2

Our theoretical and empirical results highlight the importance of Assumption 4.2. Accordingly, we present a hypothesis test to check when Assumption 4.2 holds in practice. Recall the definitions of $(\boldsymbol{H}, \boldsymbol{H}_\perp)$ and $(\boldsymbol{H}', \boldsymbol{H}_\perp')$ as defined at the start of Section 4.

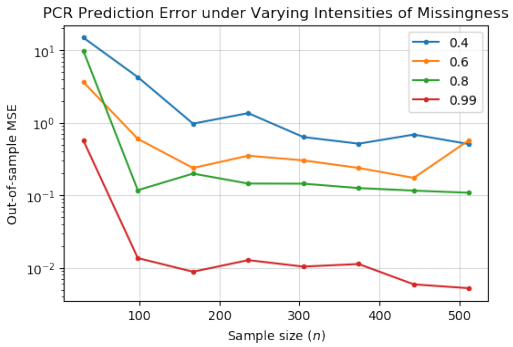

Figure 5: Plots of PCR's MSE (log scale) under varying intensities of MCAR as controlled by ρ. The MSE decays as the sample size increases for each value of ρ.

We consider the hypotheses

$$H_0: \text{ rowspan}(\boldsymbol{X}') \subseteq \text{rowspan}(\boldsymbol{X}) \quad \text{and} \quad H_1: \text{ rowspan}(\boldsymbol{X}') \nsubseteq \text{rowspan}(\boldsymbol{X}).$$

Since $(\boldsymbol{X}, \boldsymbol{X}')$ are unobserved, we use $(\boldsymbol{Z}, \boldsymbol{Z}')$ as proxies. To this end, let $\widehat{\boldsymbol{H}}^k$ and $\widehat{\boldsymbol{H}}'^{\ell}$ denote the projection matrices formed by the right singular vectors of $\widetilde{\boldsymbol{Z}}^k$ and $\widetilde{\boldsymbol{Z}}'^{\ell}$, respectively; see Section 2.2 for a recall of relevant notation. We then define our test statistic as

$$\widehat{\tau} = \|(\boldsymbol{I} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2.$$

In words, $\widehat{\tau}$ measures the ℓ_2-distance between the in- and out-of-sample covariates represented by the rowspaces of $\widetilde{\boldsymbol{Z}}^k$ and $\widetilde{\boldsymbol{Z}}'^{\ell}$, respectively.

We define the test as follows: for any significance level $\alpha \in (0, 1)$ and corresponding critical value $\tau(\alpha)$, retain H_0 if $\widehat{\tau} \leq \tau(\alpha)$ and reject H_0 if $\widehat{\tau} > \tau(\alpha)$. In Sections 6.1 and 6.2 below, we discuss two approaches to perform the hypothesis test.

6.1 A Theory-Based Approach

We first provide a theory-based approach in defining $\tau(\alpha)$. Formally, let

$$\tau(\alpha) = r' \left(\frac{C\varsigma^2 \phi^2(\alpha/2)}{s_r^2} + \frac{C\varsigma^2(\phi'(\alpha/2))^2}{(s'_{r'})^2} + \frac{C\varsigma\phi(\alpha/2)}{s_r} \right), \tag{13}$$

where $C \geq 0$ is an absolute constant, $\text{Var}(w_{ij}) \leq \varsigma^2$, $\phi(a) = \sqrt{n} + \sqrt{p} + \sqrt{\log(1/a)}$; $\phi'(a) = \sqrt{m} + \sqrt{p} + \sqrt{\log(1/a)}$; and recall that s_ℓ, s'_ℓ are the ℓ-th largest singular values of $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. See Appendix F for the derivation of (13).

6.1.1 Type I and Type II Guarantees

Given our choice of $\widehat{\tau}$ and $\tau(\alpha)$, we control both Type I and Type II errors of our test. For ease of exposition, we will consider a more restrictive form of Assumption 3.2, namely that the entries of the covariate noise are independent and $(\boldsymbol{Z}, \boldsymbol{Z}')$ are fully observed.

Theorem 6.1 *Consider Assumption 3.2 with the following conditions: (i) the entries of* $\{\boldsymbol{w}_i : i \leq n+m\}$ *are independent random variables satisfying* $\mathrm{Var}(w_{ij}) = \varsigma^2$*; (ii)* $\rho = 1$*; (iii)* $k = r$*,* $\ell = r'$*. Fix any* $\alpha \in (0,1)$*. Then there exists an absolute constant* $C \geq 0$*, depending only on the noise distribution defined in* (13)*, such that under* H_0*, the Type I error is bound by* $\mathbb{P}(\widehat{\tau} > \tau(\alpha)) \leq \alpha$*. To bound the Type II error under* H_1*, suppose further that*

$$r' > \|\boldsymbol{H}\boldsymbol{H}'\|_F^2 + 2\tau(\alpha) + \frac{C\varsigma r'\phi'(\alpha/2)}{s'_{r'}}. \tag{14}$$

Then the Type II error satisfies $\mathbb{P}(\widehat{\tau} \leq \tau(\alpha)) \leq \alpha$*.*

The particular C for which Theorem 6.1 holds depends on the underlying distribution of the covariate noise $\boldsymbol{w}_i$. C can be made explicit for certain classes of distributions; as an example, Corollary 6.1 specializes Theorem 6.1 to when $\boldsymbol{w}_i$ are normally distributed.

Corollary 6.1 *Consider the setup of Theorem 6.1 with* $C = 4$*, and suppose that* $\boldsymbol{w}_i$ *is normally distributed for all* $i \leq n+m$*. Then, under* H_0*, we have* $\mathbb{P}(\widehat{\tau} > \tau(\alpha)) \leq \alpha$*, and under* H_1*, we have* $\mathbb{P}(\widehat{\tau} \leq \tau(\alpha)) \leq \alpha$*.*

We argue (14) is not a restrictive condition. Under H_1, observe that $r' > \|\boldsymbol{H}\boldsymbol{H}'\|_F^2$ always holds. If Assumptions 4.1 and 4.3 hold, then one can easily verify that the latter two terms on the right-hand side of (14) decay to zero as (n, m, p) grow; hence, our type I and type II errors can be simultaneously minimized if our sample size increases.

6.1.2 Computing the Critical Value

Computing $\tau(\alpha)$ requires estimating (i) ς^2; (ii) r, r'; (iii) $s_r, s'_{r'}$. Under our assumptions, the covariance of $\boldsymbol{w}$ can be estimated from the sample covariance matrices of $(\boldsymbol{Z}, \boldsymbol{Z}')$. By standard random matrix theory, the singular values of $\boldsymbol{Z}$ and $\boldsymbol{X}$ are close. Thus, as discussed in Section 2.4.2, the spectrum of $\boldsymbol{Z}$ serves as a good proxy to estimate (r, s_r). Analogous arguments hold for $\boldsymbol{Z}'$ with respect to $\boldsymbol{X}'$. Corollary 6.2 specializes $\tau(\alpha)$ under Assumptions 4.1 and 4.3.

Corollary 6.2 *Let the setup of Theorem 6.1 hold. Suppose Assumptions 4.1 and 4.3 hold. Then,* $\tau(\alpha) = O\Big(\frac{\sqrt{\log(1/\alpha)}}{\min\{\sqrt{n}, \sqrt{m}, \sqrt{p}\}}\Big)$*.*

If we consider the noiseless case, $\boldsymbol{w}_i = \boldsymbol{0}$, then $\tau(\alpha) = 0$. More generally, if the spectrum of $\boldsymbol{X}$ and $\boldsymbol{X}'$ are well-balanced, then Corollary 6.2 establishes that $\tau(\alpha) = o(1)$, even in the presence of noise. We remark that Corollary 6.1 allows for exact constants in the definition of $\tau(\alpha)$ under the Gaussian noise model.

6.2 A Practical Approach

We now provide a practical approach to computing $\tau(\alpha)$. To build intuition, observe that $\widehat{\tau}$ represents the remaining spectral energy of $\boldsymbol{H}'$ not contained within $\boldsymbol{H}$. Further, we note $\widehat{\tau}$ is trivially bounded by r'. Thus, one can fix some fraction $\alpha \in (0,1)$ and reject H_0 if $\widehat{\tau} > \tau(\alpha)$, where $\tau(\alpha) = r'\alpha$. In words, if more than α fraction of the spectral energy (sum of squared singular values) of $\boldsymbol{H}'$ lies outside the span of $\boldsymbol{H}$, then the alternative

test rejects H_0. We remark that this variant is likely more robust compared to its exact computation counterpart in (13), which requires estimating several "nuisance" quantities and varies with the underlying modeling assumptions on the covariate noise and singular values. Accordingly, without knowledge of these quantities, we recommend the practical approach. To see how the practical heuristic plays out in practice, see (Squires et al., 2022, Section 7.3).

7. Synthetic Controls

To provide a concrete application of our results, we contextualize our claims in Section 4 for synthetic controls (Abadie and Gardeazabal, 2003; Abadie et al., 2010), which has emerged as a leading approach for policy evaluation with observational data (Athey and Imbens, 2017). Towards this, we connect synthetic controls to (high-dimensional) error-in-variables regression with fixed design.

7.1 Synthetic Controls Framework

Consider a panel data format where observations of $p+1$ units, indexed as $\{0, \ldots, p\}$, are collected over $n+m$ time periods. Each unit i at time t is characterized by two potential outcomes, $Y_{ti}(1)$ and $Y_{ti}(0)$, corresponding to the outcomes under treatment and absence of treatment (i.e., control), respectively (Neyman, 1923; Rubin, 1974). For each unit, we observe their potential outcomes according to their treatment status, i.e., we either observe $Y_{ti}(0)$ or $Y_{ti}(1)$, never both. Let Y_{ti} denote the observed outcome. For ease of exposition, we consider a single treated unit indexed by the zeroth unit and referred to as the target. We refer to the remaining units as the control group.

We observe all $p+1$ units under control for the first n time periods. In the remaining m time periods, we continue to observe the control group without treatment but observe the target unit with treatment. Precisely,

$$Y_{ti} = \begin{cases} Y_{ti}(0) & \text{for all } t \leq n \text{ and } i \geq 0, \\ Y_{ti}(0) & \text{for all } t > n \text{ and } i \geq 1, \\ Y_{ti}(1) & \text{for all } t > n \text{ and } i = 0. \end{cases}$$

We call the first n and final m time steps the pre- and post-treatment periods, respectively. We encode the control units' pre- and post-treatment observations into $\boldsymbol{Z} = [Y_{ti} : t \leq n, i \geq 1] \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Z}' = [Y_{ti} : t > n, i \geq 1] \in \mathbb{R}^{m \times p}$, respectively. We encode the target unit's pretreatment observations into $\boldsymbol{y} = [Y_{t0} : t \leq n] \in \mathbb{R}^n$. With these concepts in mind, we connect the synthetic controls framework to our setting of interest.

7.1.1 Out-of-Sample Prediction

Synthetic controls tackles the counterfactual question: *what would have happened to the target unit in the absence of treatment?* Formally, the goal is to estimate the (expected) counterfactual vector $\mathbb{E}[\boldsymbol{y}'(0)]$, where $\boldsymbol{y}'(0) = [Y_{t0}(0) : t > n] \in \mathbb{R}^m$. Methodologically, this is answered by regressing $\boldsymbol{y}$ on $\boldsymbol{Z}$ and applying the regression coefficients $\widehat{\boldsymbol{\beta}}$ to $\boldsymbol{Z}'$ to estimate the treated unit's expected potential outcomes under control during the post-treatment

period. From this perspective, we identify that counterfactual estimation is precisely out-of-sample prediction.

7.1.2 Error-in-Variables

As is typical in panel studies, potential outcomes are modeled as the addition of a latent factor model and a random variable in order to model measurement error and/or mis-specification (Abadie, 2021). That is, $Y_{ti}(0) = \langle \boldsymbol{u}_t, \boldsymbol{v}_i \rangle + \varepsilon_{ti}$, where $\boldsymbol{u}_t, \boldsymbol{v}_i \in \mathbb{R}^r$ represent latent time and unit features with r much smaller than (n, m, p), and $\varepsilon_{ti} \in \mathbb{R}$ models the stochasticity. This is also known as an interactive fixed effects model (Bai, 2009). Put differently, the observed matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ are viewed as noisy instantiations of $\boldsymbol{X} = \mathbb{E}[\boldsymbol{Z}]$ and $\boldsymbol{X}' = \mathbb{E}[\boldsymbol{Z}']$, where $\boldsymbol{X}, \boldsymbol{X}'$ are low-rank matrices. They represent the matrices of latent expected potential outcomes, which are a function of the latent time and unit factors. Since $\widehat{\boldsymbol{\beta}}$ is learned using $\boldsymbol{Z}$ not $\boldsymbol{X}$, synthetic controls is an instance of error-in-variables regression.

Remark 1 (Clarifying MCAR entries) *As described in Section 3, we require the entries in $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ to be missing completely at random (MCAR). We emphasize that these missing elements do not correspond to our counterfactual estimands. Traditionally, the SC setup assumes $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ are fully observed, i.e., the canonical SC setting does not even consider missing data for the control units. In this view, our estimator and analysis allows for a more general observation pattern than that typically studied in the SC literature. However, readers who find the MCAR setting to be implausible can proceed with the balanced panel data setting in mind. For recent progress on extending the SC setup to include more general observation patterns, please see Ben-Michael et al. (2021) and references therein.*

7.1.3 Linear Model

The underlying premise behind synthetic controls is that the target unit is a weighted composition of control units. In our setup, this translates more formally as the existence of a linear model $\boldsymbol{\beta}^* \in \mathbb{R}^p$ satisfying

$$\mathbb{E}[Y_{t0}(0)] = \sum_{i \geq 1} \beta_i^* \mathbb{E}[Y_{ti}(0)] \implies Y_{t0}(0) = \sum_{i \geq 1} \beta_i^* \mathbb{E}[Y_{ti}(0)] + \varepsilon_{t0}$$

for every $t \in [n + m]$, i.e., $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = [\varepsilon_{t0} : t \leq n]$. We note that (Agarwal et al., 2021) establish that such a $\boldsymbol{\beta}^*$ exists w.h.p. if r is much smaller than (n, m, p).

7.1.4 Fixed Design

Several works in the literature, e.g., Agarwal et al. (2021), enforce the latent time factors to be sampled i.i.d. Subsequently, the pre- and post-treatment data under control are also i.i.d. In contrast, we consider a fixed design setting that avoids distributional assumptions on the expected potential outcomes. This allows us to model settings with underlying time trends or shifting ideologies, which are likely present in many panel studies.

7.2 Novel Guarantees for the Synthetic Controls Literature

With our connection established, we transfer our theoretical results to the synthetic controls framework. In particular, we analyze the robust synthetic controls (RSC) estimator of Amjad et al. (2018) and its extension in Amjad et al. (2019), which learns $\widehat{\boldsymbol{\beta}}$ via PCR.

7.2.1 Model Identification

Intuitively, $\boldsymbol{\beta}^*$ defines the synthetic control group. That is, the magnitude (and sign) of the ith entry specifies the contribution of the ith control unit in the construction of the target unit. Theorem 4.1 establishes that RSC consistently identifies the unique synthetic control group with minimum ℓ_2-norm.

7.2.2 Counterfactual Estimation

We denote RSC's estimate of the expected counterfactual trajectory as $\widehat{\mathbb{E}}[\boldsymbol{y}'(0)] = [\widehat{\mathbb{E}}[Y_{t0}(0)] : t > n]$. The counterfactual estimation error is then

$$\frac{1}{m}\|\widehat{\mathbb{E}}[\boldsymbol{y}'(0)] - \mathbb{E}[\boldsymbol{y}'(0)]\|_2^2, \tag{15}$$

which precisely corresponds to (9). Theorem 4.2 immediately leads to a vanishing bound on (15) as (n, m, p) grow. The exact finite-sample rates given in Theorem 4.2 improve upon the best known rate provided in Agarwal et al. (2021), which is only established in expectation and for random designs.

7.3 Examining Assumption 4.2 for Two Synthetic Controls Studies

We revisit two canonical synthetic controls case studies: (i) terrorism in Basque Country (Abadie and Gardeazabal, 2003) and (ii) California's Proposition 99 (Abadie et al., 2010). These studies have been used extensively to explain the utility of the synthetic controls method. We apply the practical variant of our hypothesis test for Assumption 4.2 in Section 6.2 to study the potential feasibility of counterfactual inference in both studies.

7.3.1 Terrorism in Basque Country

Background & setup. Our first study evaluates the economic ramifications of terrorism on the Basque Country of Spain. Our data is comprised of the per-capita GDP associated with 17 Spanish regions over 43 years. Basque Country is the sole treated unit that is affected by terrorism; the remaining $p = 16$ regions are the control regions that are relatively unaffected by terrorism. The pre- and post-intervention durations are $n = 14$ and $m = 29$ years, respectively. We note that the original work of Abadie and Gardeazabal (2003) uses 13 additional predictor variables for each region, including demographic information pertaining to one's educational status, and average shares for six industrial sectors. We only use information related to the outcome of interest, i.e., the per-capita GDP.

Hypothesis test results. We consider $\alpha = 0.05$. We estimate $r' = 3$ via the universal data-driven approach of Gavish and Donoho (2014). This sets $\tau(\alpha) = 0.15$. Estimating r analogously to r', we obtain $r = 5$ and $\widehat{\tau} = 0.61$. Since $\widehat{\tau} > \tau(\alpha)$, our test suggests that the PCR-based method of Amjad et al. (2018) may not be suitable for this study under

our assumptions. In fact, our test only passes for (effectively) $\alpha > 0.21$, which roughly translates to allowing for over 21% of the spectral energy of $\boldsymbol{H}'$ to fall outside of $\boldsymbol{H}$.

7.3.2 California Proposition 99

Background & setup. Our second study evaluates the effect of California's Proposition 99 on the consumption of tobacco. Our data is comprised of annual per-capita cigarette sales at the state level for 39 U.S. states for 31 years. With the exception of California, the other states in this study neither adopted an anti-tobacco program or raised cigarette sales taxes by 50 cents or more. As such, the remaining $p = 38$ states are considered the control states and California is considered the treated state. The pre- and post-intervention durations are $n = 18$ and $m = 13$ years, respectively. The original work of Abadie et al. (2010) uses six additional covariates per state. We do not include these variables in our study.

Hypothesis test results. We consider $\alpha = 0.05$. Estimating (r, r') as above, we obtain $r = 4$ and $r' = 3$, which yields $\tau(\alpha) = 0.15$ and $\widehat{\tau} = 1.63$. Again, we have $\widehat{\tau} > \tau(\alpha)$, which suggests that PCR-based methods may be ill-suited to produce reliable counterfactual estimates under our assumptions. Our test, therefore, only passes for (effectively) $\alpha > 0.55$.

7.3.3 Discussion of Findings

Although our tests do not pass for either study, our results are not meant to discredit the previous conclusions drawn in Amjad et al. (2018) and Agarwal et al. (2021). Rather, our tests highlight that these studies warrant further investigation. We hope our findings not only motivate the usage of this test, but also spark the development of new robustness tests to stress investigate the assumptions that underlie statistical methods and thus the associated causal conclusions drawn from these methods.

8. Related works

This section discusses related prior works from several literatures.

8.1 Principal Component Regression

Since its introduction in Jolliffe (1982), there have been several notable works analyzing PCR, including Bair et al. (2006); Agarwal et al. (2021); Chao et al. (2019). We pay particular attention to Agarwal et al. (2021) given their closeness to this article.

8.1.1 Model Identification

Agarwal et al. (2021) purely focuses on prediction and thus, do not provide any results for model identification. This work proves that PCR identifies the unique minimum ℓ_2-norm model with non-asymptotic rates of convergence.

8.1.2 Out-of-Sample Prediction

Agarwal et al. (2021) shows that PCR's out-of-sample prediction error decays as $\widetilde{O}(1/\sqrt{n})$ when $m, p = \Theta(n)$. Agarwal et al. (2021) conjecture that their "slow" rate is an artefact of

their Rademacher complexity arguments. By leveraging our model identification result in Theorem 4.1, we establish the "fast" rate of $\widetilde{O}(1/n)$.

8.1.3 Framework

Learning setup. Agarwal et al. (2021) considers a transductive learning setting, where *both* the in- and out-of-sample covariates are accessible upfront. This work, in comparison, considers the classical supervised learning setup, where the out-of-sample covariates are not revealed during training.

Covariate design. Agarwal et al. (2021) considers a random design setting with i.i.d. covariates. By contrast, we consider a fixed design setting. As Shao and Deng (2012) notes, estimation in high-dimensional regimes with fixed designs is very different from those with random designs due to the identifiability of the model parameter. Additionally, since we treat the covariates as deterministic, we do *not* impose that the in- and out-of-sample covariates obey the same distribution. Under the linear algebraic condition of Assumption 4.2, we prove that PCR achieves consistent out-of-sample prediction in Corollary 4.2.

8.2 Functional Principal Component Analysis

We consider functional principal component analysis (fPCA), which generalizes PCA to infinite-dimensional operators (Yao et al., 2005; Hall et al., 2006; Li and Hsing, 2010; Descary et al., 2019). This literature often assumes access to n randomly sampled trajectories at p locations, which are carefully chosen from a grid with minor perturbations, forming an $n \times p$ data matrix, $\boldsymbol{D}$. Thus, $\boldsymbol{D}^\top \boldsymbol{D}$ is the empirical proxy of the underlying covariance kernel that corresponds to these random trajectories. Under appropriate assumptions on the trajectories, the $\boldsymbol{D}^\top \boldsymbol{D}$ matrix can be represented as the additive sum of a low-rank matrix with a noise matrix. This resembles the low-rank matrix estimation problem with a key difference being that *all* entries here are fully observed. In Descary et al. (2019), the low-rank component is estimated by performing an explicit rank minimization, which is known to be computationally hard. The functional (or trajectory) approximation from this low-rank estimation is obtained by smoothing (or interpolation)—this is where the careful choice of locations in a grid plays an important role. The estimation error is provided with respect to the normalized Frobenius norm (i.e., Hilbert-Schmidt norm when discretized). Finally, we remark that the fPCA literature has thus far considered diverging n with fixed p or $n \gg p$.

In comparison, PCR utilizes hard singular value thresholding (HSVT), a popular method in the matrix estimation toolkit, to recover the low-rank matrix; such an approach is computationally efficient and even yields a closed form solution. As shown in Agarwal et al. (2021), PCR can be equivalently interpreted as HSVT followed by ordinary least squares. Hence, unlike the standard fPCA setup, PCR allows for missing values in the covariate matrix since HSVT recovers the underlying matrix in the presence of noisy and missing entries. Analytically, our model identification and prediction error guarantees rely on matrix recovery bounds with respect to the $\ell_{2,\infty}$-norm, which is stronger than the Frobenius norm, i.e., $(np)^{-1/2}\|\boldsymbol{A}\|_F \leq n^{-1/2}\|\boldsymbol{A}\|_{2,\infty}$. Put differently, the typical Frobenius norm bound is insufficient to provide guarantees for PCR with error-in-variables. Finally, our setting allows for both $n \ll p$ and $n \gg p$; the current fPCA literature only allows for $n \gg p$.

In this view, our work offers several directions for research within the fPCA literature: (i) allow the sampling locations to be different across the n measurements, provided there is sufficient overlap; (ii) consider settings beyond $n \gg p$; (iii) extend fPCA guarantees for computationally efficient methods like HSVT.

There has also been work on functional principal component regression (fPCR), which allows $\boldsymbol{\beta}^*$ to be an infinite-dimensional parameter. Notable works include Hall and Horowitz (2007) and Cai and Hall (2006), which consider the problems of model identification and prediction error, respectively. These works, however, do *not* allow for error-in-variables. As noted above, model identification and out-of-sample guarantees at the fast rate of $\widetilde{O}(1/n)$ for PCR with error-in-variables in the finite-dimensional case has remained elusive. Extending these results for fPCR with error-in-variables remains interesting future work.

8.3 Error-in-Variables

There are numerous prominent works in the high-dimensional error-in-variables literature, including Rosenbaum and Tsybakov (2010, 2013); Chen and Caramanis (2012, 2013); Loh and Wainwright (2012); Kaul and Koul (2015); Belloni et al. (2017a,b); Datta and Zou (2017). Below, we highlight a few key points of comparison.

8.3.1 Out-of-Sample Prediction

By and large, this literature has focused on model identification. Accordingly, the algorithms in the works above are ill-equipped to produce reliable predictions given corrupted and partially observed out-of-sample covariates. Therefore, even if the true model parameter $\boldsymbol{\beta}^*$ is known, it is unclear how prior results can be extended to establish generalization error bounds. This work shows PCR can be easily adapted to handle these cases.

8.3.2 Knowledge of Noise Distribution

Many existing algorithms explicitly utilize knowledge of the underlying noise distribution to recover $\boldsymbol{\beta}^*$. Typically, these algorithms perform corrections of the form $\boldsymbol{Z}^\top \boldsymbol{Z} - \mathbb{E}[\boldsymbol{W}^\top \boldsymbol{W}]$. To carry out this computation, one must assume access to either oracle knowledge of $\mathbb{E}[\boldsymbol{W}^\top \boldsymbol{W}]$ or obtain a good data-driven estimator for it. As Chen and Caramanis (2013) note, such an estimator can be costly or simply infeasible in many practical settings. PCR does not require any such knowledge. Instead, the PCA subroutine within PCR *implicitly* de-noises the covariates. The trade-off is that our results only hold if the number of retained singular components k is chosen to be the rank of $\boldsymbol{X}$. Although there are numerous heuristics to aptly choose k, we leave a formal analysis of PCR when k is misspecified as important future work.

8.3.3 Operating Assumptions

We compare our primary assumptions with canonical assumptions in the literature.

I: Low-rank vis-á-vis sparsity. The most popularly endowed structure in high-dimensional regression is that the model parameter $\boldsymbol{\beta}^*$ is r-sparse. This work posits that the in-sample covariate matrix $\boldsymbol{X}$ is described by r nonzero singular values. These two notions are related. If $\text{rank}(\boldsymbol{X}) = r$, then there exists an r-sparse $\tilde{\boldsymbol{\beta}}$ such that $\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$; see Proposition 3.4

of Agarwal et al. (2021). Meanwhile, if $\boldsymbol{\beta}^*$ is r-sparse, then there exists a $\tilde{\boldsymbol{X}}$ of rank r that also provides equivalent responses. In this view, the two perspectives are complementary.

With that said, it is difficult to verify the sparsity of $\boldsymbol{\beta}^*$, but the low-rank assumption on $\boldsymbol{X}$ can be examined through the singular values of $\boldsymbol{Z}$, as described in Section 2.4.2. It is also well-established that (approximately) low-rank matrices are abundant in real-world data science applications (Xu, 2018; Udell and Townsend, 2017, 2018).

II: Well-balanced spectra vis-á-vis restricted eigenvalue condition. The second common condition in the literature captures the amount of "information spread" across the rows and columns of $\boldsymbol{X}$, which leads to a bound on its smallest singular value. This is referred to as the restricted eigenvalue condition (see Definitions 1 and 2 in Loh and Wainwright (2012)), which is imposed on the empirical estimate of the covariance of $\boldsymbol{X}$. This work assumes the spectrum of $\boldsymbol{X}$ is well-balanced (Assumption 4.1). This assumption is *not* necessary for consistent estimation. Rather, it is one condition that yields a reasonable snr, which guarantees both model identification *and* vanishing out-of-sample prediction errors.

In many prior works, the restricted eigenvalue condition (and its variants) are shown to hold w.h.p. if the rows of $\boldsymbol{X}$ are i.i.d. (or at least, independent) samples from a mean zero sub-gaussian distribution. This data generating process implies that the smallest and largest singular values of $\boldsymbol{X}$ are of order $\widetilde{O}(\sqrt{n} + \sqrt{p})$. However, under the assumptions $\text{rank}(\boldsymbol{X}) = r$ and $|X_{ij}| = \Theta(1)$, one can verify that $\|\boldsymbol{X}\|_2 = \Omega(\sqrt{np/r})$. The difference in the typical magnitude of the largest singular value reflects the difference in applications in which a restricted eigenvalue assumption versus a low-rank assumption is likely to hold. The restricted eigenvalue assumption is particularly suited in applications such as compressed sensing where researchers *design* $\boldsymbol{X}$. The applications arising in the social or life sciences primarily involve observational data. In such settings, a low-rank assumption on $\boldsymbol{X}$ is arguably more suitable to capture the latent structure amongst the covariates. Ultimately, the Assumption 4.1 is similar to the restricted eigenvalue condition in that it requires the smallest and largest nonzero singular values of $\boldsymbol{X}$ to be of the same order.

It turns out that analogous assumptions are pervasive across many fields. Within the econometrics factor model literature, it is standard to assume that the factor structure is separated from the idiosyncratic errors, e.g., Assumption A of Bai and Ng (2021); within the robust covariance estimation literature, this assumption is closely related to the notion of pervasiveness, e.g., Proposition 3.2 of Fan et al. (2018); within the matrix/tensor completion literature, it is assumed that the nonzero singular values are of the same order to achieve minimax optimal rates, e.g., Cai et al. (2019). Assumption 4.1 has also been shown to hold w.h.p. for the embedded Gaussians model, which is a canonical probabilistic generating process used to analyze probabilistic PCA (Tipping and Bishop, 1999; Bishop, 1999; Agarwal et al., 2021). Finally, like the low-rank assumption, a practical benefit of the well-balanced spectra assumption is that it can be empirically examined via the same procedure outlined in Section 2.4.2.

8.4 Linear Regression with Hidden Confounding

The problem of high-dimensional error-in-variables regression is related to linear regression with hidden confounding, a common model within the causal inference and econometrics literatures (Guo et al., 2022; Ćevid et al., 2020). As noted by Guo et al. (2022), a partic-

ular class of error-in-variables models can be reformulated as linear regression with hidden confounding. Using our notation, they consider a high-dimensional model where the rows of $\boldsymbol{X}$ are sampled i.i.d. As such, $\boldsymbol{X}$ can be full-rank, but $\boldsymbol{W}$ is assumed to have low-rank structure. The aim of this work is to estimate a sparse $\boldsymbol{\beta}^*$. In comparison, we place the low-rank assumption on $\boldsymbol{X}$, and assume the rows of $\boldsymbol{W}$ are sampled independently and thus, can be full-rank. Notably, for this setup, Ćevid et al. (2020) "deconfounds" the observed covariates $\boldsymbol{Z}$ by a spectral transformation of its singular values. It is interesting future work to analyze PCR for this important and closely related scenario.

9. Conclusion

In this article, we analyze the classical method of PCR within the high-dimensional EiV framework with fixed design. We first prove that PCR consistently identifies the projection of the model parameter onto the subspace spanned by the underlying covariates, $\boldsymbol{X}$. Equipped with this result, we establish non-asymptotic out-of-sample prediction guarantees that improve upon the best known rates for PCR. Moreover, we furnish a data-driven hypothesis test to empirically assess a critical linear algebraic condition that enables PCR's vanishing prediction error. We complement our statistical claims with illustrative simulation studies. Finally, we provide a concrete application of our statistical claims within the context of the synthetic controls paradigm, thereby providing theoretical guarantees for several PCR based synthetic controls estimators.

We envision several directions for future research. One immediate next step is to establish bounds when $\boldsymbol{X}$ is only approximately low-rank. Within this context, our analysis suggests PCR induces an additional error of the form $\|(\boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{V}_r^\top)\tilde{\boldsymbol{\beta}}^*\|_2$, where $\boldsymbol{V}_r$ is formed from the top r principal components of $\boldsymbol{X}$. This is the unavoidable model misspecification error that results from taking a rank r approximation of $\boldsymbol{X}$. It stands to reason that *soft* singular value thresholding (SVT), which appropriately down-weights the singular values of $\widetilde{\boldsymbol{Z}}$, may be a more appropriate algorithmic approach as opposed to the *hard* SVT.

Another future line of research is to bridge our out-of-sample prediction analysis with the analysis of over-parameterized estimators. Bartlett et al. (2020), for instance, demonstrates that the minimum ℓ_2-norm linear regression solution predicts well out-of-sample despite a perfect fit to noisy in-sample data; this phenomena is known as *benign overfitting*. To establish their result, Bartlett et al. (2020) introduces two notions of "effective rank" of the data covariance, and characterize linear regression problems that exhibit benign overfitting with respect to these quantities. In comparison, this work characterizes the out-of-sample prediction performance of PCR with respect to the ℓ_2-distance between the in- and out-of-sample covariates (see Assumption 4.2). Accordingly, one exciting research agenda is to explore the interplay of these two conceptions for over-parameterized linear estimators. This may also have implications for approximately low-rank settings. On a related note, developing Monte-Carlo or resampling-based tests, as an alternative to the approaches in Sections 6.1 and 6.2, to empirically assess Assumption 4.2 is a worthwhile exploration.

Within the context of synthetic controls, it would also be interesting to innovate upon PCR by including the simplex constraint that is often utilized in synthetic controls applications. Such a constrained variant of PCR offers greater interpretability and is consistent

with the spirit of the original synthetic controls formulation. Other constraints such as a bias towards a sparse model parameter may be of similar interest.

Finally, a critical move forward is to rigorously quantify the uncertainty in PCR's parameter and prediction estimates to enable valid inference. As an important byproduct, such an innovation would also immediately provide uncertainty estimates for PCR based synthetic controls estimators, which have remained elusive in the literature, and contribute to the exciting progress made on uncertainty quantification for synthetic controls more generally as in Matias D. Cattaneo and Titiunik (2021); Cattaneo et al. (2024) and Shaikh and Toulis (2021).

Acknowledgments

We thank Peng Ding and various members within MIT's Laboratory for Information and Decision Systems (LIDS) for useful discussions and guidance. The data and code to reproduce the results in this article are available at https://github.com/deshen24/principal-component-regression.

Supplementary Materials

The supplementary material is structured as follows. Appendix A details the setups for the simulations presented in Section 5. Appendix B presents an additional simulation that studies PCR's prediction accuracy when k is misspecified. Appendices C and D prove Theorems 4.1 and 4.2, respectively. Appendix E contains helpful lemmas that assist in the proofs of Theorems 4.1 and 4.2. Appendix F proves Theorem 6.1 and Corollary 6.1. Appendix G expounds on the discussion in Section 4.3.3 on providing a lower bound on PCR's parameter estimation error.

Appendix A. Illustrative Simulations: Details

We present the generative models in our simulation studies in Section 5.

A.1 PCR Identifies the Minimum ℓ_2-norm Model

We generate $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$, where the entries of $\boldsymbol{U}, \boldsymbol{V}$ are sampled independently from a standard normal distribution. Next, we generate $\boldsymbol{\beta}^* \in \mathbb{R}^p$ by sampling from a multivariate standard normal vector with independent entries, and normalize it by $\|\boldsymbol{\beta}^*\|_2$ so that it has unit norm. We define $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{X}^\dagger \boldsymbol{X}\boldsymbol{\beta}^*$. For each simulation repeat, we independently sample the entries of $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ from a normal distribution with mean 0 and variance $\sigma^2 = 0.2$. The entries of $\boldsymbol{W} \in \mathbb{R}^{n \times p}$ are sampled in an identical fashion. We then define our observed response vector as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and observed covariate matrix as $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$. For simplicity, we do not mask any of the entries.

A.2 PCR is Robust to Covariate Shifts

We generate $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$ as in Appendix A.1. Next, we generate four different out-of-sample covariates, defined as $\boldsymbol{X}_1', \boldsymbol{X}_2', \boldsymbol{X}_3', \boldsymbol{X}_4'$ via the following procedure: We independently sample the entries of $\boldsymbol{U'}_1$ from a standard normal distribution, and define $\boldsymbol{X}_1' = \boldsymbol{U'}_1\boldsymbol{V}^\top$. We define $\boldsymbol{X}_2' = \boldsymbol{U'}_2\boldsymbol{V}^\top$ similarly with the entries of $\boldsymbol{U'}_2$ sampled from $\mathcal{N}(0, 5)$. Next, we independently sample the entries of $\boldsymbol{U'}_3$ from Uniform$[-\sqrt{3}, \sqrt{3}]$, and define $\boldsymbol{X}_3' = \boldsymbol{U'}_3\boldsymbol{V}^\top$. We define $\boldsymbol{X}_4' = \boldsymbol{U'}_4\boldsymbol{V}^\top$ similarly with the entries of $\boldsymbol{U'}_4$ sampled from Uniform$[-\sqrt{15}, \sqrt{15}]$.

By construction, the mean and variance of the entries in $\boldsymbol{X}_3'$ match that of $\boldsymbol{X}_1'$; an analogous relationship holds between $\boldsymbol{X}_4'$ and $\boldsymbol{X}_2'$. While $\boldsymbol{X}_1'$ follows the same distribution as that of $\boldsymbol{X}$, there is a clear distribution shift from $\boldsymbol{X}$ to $\boldsymbol{X}_3', \boldsymbol{X}_2', \boldsymbol{X}_4'$.

We proceed to generate $\boldsymbol{\beta}^*$ from a standard multivariate normal. We define $\boldsymbol{\theta}_1' = \boldsymbol{X}_1'\boldsymbol{\beta}^*$, and define $\boldsymbol{\theta}_2', \boldsymbol{\theta}_3', \boldsymbol{\theta}_4'$ analogously. Further, the entries of $\boldsymbol{\varepsilon}$ and $\boldsymbol{W}, \boldsymbol{W}'$ are independently sampled from a normal distribution with variance $\sigma^2 = 0.2$. We define the training responses as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and observed training covariates as $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$. The first set of observed testing covariates is defined as $\boldsymbol{Z}_1' = \boldsymbol{X}_1' + \boldsymbol{W}'$, with analogous definitions for $\boldsymbol{Z}_2', \boldsymbol{Z}_3', \boldsymbol{Z}_4'$.

A.3 PCR Generalizes under Assumption 4.2

We generate $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$ as in Appendix A.1. We now generate two different testing covariates. First, we generate $\boldsymbol{X}_1' = \boldsymbol{U'}\boldsymbol{V}^\top$, where the entries of $\boldsymbol{U'}$ are independently

sampled from a normal distribution with mean zero and variance 5. As such, it follows that Assumption 4.2 immediately holds between $\boldsymbol{X}_1'$ and $\boldsymbol{X}$, though they do not obey the same distribution. Next, we generate $\boldsymbol{X}_2' = \boldsymbol{U}\boldsymbol{V}'^T$, where the entries of $\boldsymbol{V}'$ are independently sampled from a standard normal (just as in $\boldsymbol{V}$). In doing so, we ensure that $\boldsymbol{X}_2'$ and $\boldsymbol{X}$ follow the same distribution, though Assumption 4.2 no longer holds.

We generate $\boldsymbol{\beta}^*$ as in Appendix A.2, and define $\boldsymbol{\theta}_1' = \boldsymbol{X}_1'\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}_2' = \boldsymbol{X}_2'\boldsymbol{\beta}^*$. We also generate $\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{W}'$ as in Appendix A.2. In turn, we define the training data as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, and testing data as $\boldsymbol{Z}_1' = \boldsymbol{X}_1' + \boldsymbol{W}'$ and $\boldsymbol{Z}_2' = \boldsymbol{X}_2' + \boldsymbol{W}'$.

A.4 PCR Generalizes with MCAR Entries

We generate $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$ as in Appendix A.1 and generate $\boldsymbol{X}' = \boldsymbol{U}'\boldsymbol{V}^\top$, where the entries of $\boldsymbol{U}'$ are independently sampled from a standard normal. As such, it follows that Assumption 4.2 immediately holds between $\boldsymbol{X}'$ and $\boldsymbol{X}$.

We generate $\boldsymbol{\beta}^*$ as in Appendix A.2, and define $\boldsymbol{\theta}' = \boldsymbol{X}'\boldsymbol{\beta}^*$. We also generate $(\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{W}')$ as in Appendix A.2. There are ρ entries in $\boldsymbol{\Pi}, \boldsymbol{\Pi}'$ that are randomly assigned the value 1, and each iteration considers a different permutation of revealed entries. Putting everything together, we define the training data as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and $\boldsymbol{Z} = (\boldsymbol{X} + \boldsymbol{W}) \circ \boldsymbol{\Pi}$, and testing data as $\boldsymbol{Z}' = (\boldsymbol{X}' + \boldsymbol{W}') \circ \boldsymbol{\Pi}'$.

Appendix B. Simulation: PCR with a Misspecified Choice of k

We present an additional simulation to complement those presented in Section 5. The purpose of this simulation is to study PCR's prediction accuracy when the number of principal components k is misspecified. Specifically, we compare PCR's generalization error under two settings: (i) $k > r = \text{rank}(\boldsymbol{X})$ and (ii) $k < r = \text{rank}(\boldsymbol{X})$.

Setup. We choose the same relative scalings of (n, m, p, r) and generate $\boldsymbol{\beta}^*$ as in Section 5.2. For each n, we generate $\boldsymbol{X} \sim \mathcal{D}_1$. We then generate the out-of-sample covariates as $\boldsymbol{X}' \sim \mathcal{D}_2$ with $\mathcal{D}_2 \neq \mathcal{D}_1$ that obeys Assumption 4.2. Next, we define $\boldsymbol{\theta}' = \boldsymbol{X}'\boldsymbol{\beta}^*$. We conduct 100 simulation repeats. For each repeat, we sample $(\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{W}')$ to construct $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, and $\boldsymbol{Z}' = \boldsymbol{X}' + \boldsymbol{W}'$.

Results. For each simulation repeat, we apply PCR on $(\boldsymbol{y}, \boldsymbol{Z})$ twice: (i) $\widehat{\boldsymbol{\beta}}_>$ with $k = r+5 > r$ and (ii) $\widehat{\boldsymbol{\beta}}_<$ with $k = r - 5 < r$. We then define $\widehat{\boldsymbol{y}}_>'$ from the de-noised version of $\boldsymbol{Z}'$ and $\widehat{\boldsymbol{\beta}}_>$; define $\widehat{\boldsymbol{y}}_<'$ analogously with $\widehat{\boldsymbol{\beta}}_<$. Figure 6 displays the MSEs of $\widehat{\boldsymbol{y}}_>'$ and $\widehat{\boldsymbol{y}}_<'$ with respect to $\boldsymbol{\theta}'$. When $k > r$, the MSE degrades gracefully as the sample size increases; by contrast, when $k < r$, the MSE is stagnant across varying sample sizes. Our findings suggest that practitioners should err on the side of choosing more principal components than less.

Appendix C. Proof of Theorem 4.1

We start with some useful notation. Note $\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*$. Let $\boldsymbol{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}^* + \boldsymbol{\varepsilon}$ be the vector notation of (1) with $\boldsymbol{y} = [y_i : i \leq n] \in \mathbb{R}^n$, $\boldsymbol{\varepsilon} = [\varepsilon_i : i \leq n] \in \mathbb{R}^n$. Throughout, let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ denote the singular value decomposition (SVD) of $\boldsymbol{X}$. Recall that we write

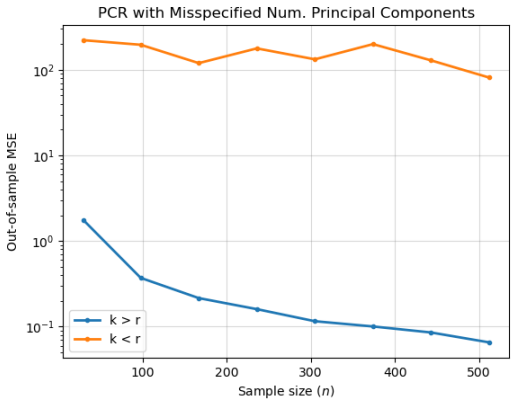

Figure 6: Plots of PCR's MSE (log scale) under two cases: (i) $k > r$ (blue); (ii) $k < r$ (orange). Case (i) achieves a diminishing MSE while case (ii) suffers from non-vanishing MSE.

$\widetilde{\boldsymbol{Z}} = \widehat{\rho}^{-1}\boldsymbol{Z} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{V}}^\top$ for the SVD of $\widetilde{\boldsymbol{Z}}$. Its truncation using the top k singular components is denoted as $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{S}}_k\widehat{\boldsymbol{V}}_k^\top$.

Further, we will often use the following bound: for any $\boldsymbol{A} \in \mathbb{R}^{a\times b}$, $\boldsymbol{v} \in \mathbb{R}^b$,

$$\|\boldsymbol{A}\boldsymbol{v}\|_2 = \|\sum_{j=1}^{b} \boldsymbol{A}_{\cdot j} v_j\|_2 \le \big(\max_{j\le b} \|\boldsymbol{A}_{\cdot j}\|_2\big)\big(\sum_{j=1}^{b} |v_j|\big) = \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{v}\|_1, \tag{S1}$$

where $\|\boldsymbol{A}\|_{2,\infty} = \max_j \|\boldsymbol{A}_{\cdot j}\|_2$ with $\boldsymbol{A}_{\cdot j}$ representing the j-th column of $\boldsymbol{A}$.

As discussed in Section 4.1, we will consider $\tilde{\boldsymbol{\beta}}^*$ as our model parameter of interest. This corresponds to the unique minimum ℓ_2-norm model parameter satisfying (1) for $i \le n$. As a result, it follows that

$$\boldsymbol{V}_\perp^\top \tilde{\boldsymbol{\beta}}^* = \boldsymbol{0}, \tag{S2}$$

where $\boldsymbol{V}_\perp$ represents a matrix of orthornormal basis vectors that span the nullspace of $\boldsymbol{X}$.

Similarly, let $\widehat{\boldsymbol{V}}_{k,\perp} \in \mathbb{R}^{p\times(p-k)}$ be a matrix of orthonormal basis vectors that span the nullspace of $\widetilde{\boldsymbol{Z}}^k$; thus, $\widehat{\boldsymbol{V}}_{k,\perp}$ is orthogonal to $\widehat{\boldsymbol{V}}_k$. Then,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 &= \|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*) + \widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\
&= \|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \;+\; \|\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\
&= \|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \;+\; \|\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top\tilde{\boldsymbol{\beta}}^*\|_2^2.
\end{aligned} \tag{S3}$$

Note that in the last equality we have used Property 2.1, which states that $\widehat{\boldsymbol{V}}_{k,\perp}^\top\widehat{\boldsymbol{\beta}} = \boldsymbol{0}$. Next, we bound the two terms in (S3).

Bounding $\|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2$. To begin, note that

$$\|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 = \|\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2, \tag{S4}$$

since $\widehat{\boldsymbol{V}}_k$ has orthonormal columns. Next, consider

$$\|\widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \le 2\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2^2 + 2\|\boldsymbol{X}\tilde{\boldsymbol{\beta}}^* - \widetilde{\boldsymbol{Z}}^k\tilde{\boldsymbol{\beta}}^*\|_2^2$$

$$\leq 2\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2^2 + 2\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2,$$

where we used (S1). Recall that $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{S}}_k\widehat{\boldsymbol{V}}_k^\top$. Therefore,

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 &= (\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)^\top\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{S}}_k^2\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\\ &= (\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*))^\top\widehat{\boldsymbol{S}}_k^2(\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*))\\ &\geq \widehat{s}_k^2\|\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2.\end{aligned}$$

Therefore using (S4), we conclude that

$$\|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \leq \frac{2}{\widehat{s}_k^2}\Big(\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2^2 + \|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2\Big). \tag{S5}$$

Next, we bound $\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2$.

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{y}\|_2^2 &= \|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^* - \boldsymbol{\varepsilon}\|_2^2\\ &= \|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 - 2\langle\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle.\end{aligned} \tag{S6}$$

By Property 2.1 we have,

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{y}\|_2^2 \leq \|\widetilde{\boldsymbol{Z}}^k\tilde{\boldsymbol{\beta}}^* - \boldsymbol{y}\|_2^2 \;&=\; \|(\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\tilde{\boldsymbol{\beta}}^* - \boldsymbol{\varepsilon}\|_2^2\\ &= \|(\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\tilde{\boldsymbol{\beta}}^*\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2 - 2\langle(\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle.\end{aligned} \tag{S7}$$

From (S6) and (S7), we have

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}^k\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2^2 &\leq \|(\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\tilde{\boldsymbol{\beta}}^*\|_2^2 + 2\langle\widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\\ &\leq \|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + 2\langle\widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle,\end{aligned} \tag{S8}$$

where we used (S1). From (S5) and (S8), we conclude that

$$\|\widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \leq \frac{4}{\widehat{s}_k^2}\Big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\Big). \tag{S9}$$

Bounding $\|\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top\tilde{\boldsymbol{\beta}}^*\|_2^2$. Consider

$$\begin{aligned}\|\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top\tilde{\boldsymbol{\beta}}^*\|_2 &= \|(\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\top)\tilde{\boldsymbol{\beta}}^* + \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\top\tilde{\boldsymbol{\beta}}^*\|_2\\ &\overset{(a)}{=} \|(\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\top)\tilde{\boldsymbol{\beta}}^*\|_2\\ &\leq \|\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\top\|_2\;\|\tilde{\boldsymbol{\beta}}^*\|_2,\end{aligned} \tag{S10}$$

where (a) follows from $\boldsymbol{V}_\perp^\top\tilde{\boldsymbol{\beta}}^* = \boldsymbol{0}$ due to (S2). Then,

$$\begin{aligned}\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\top &= (\mathbf{I} - \boldsymbol{V}_\perp\boldsymbol{V}_\perp^\top) - (\mathbf{I} - \widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top)\\ &= \boldsymbol{V}\boldsymbol{V}^\top - \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top.\end{aligned} \tag{S11}$$

From (S10) and (S11), it follows that

$$\|\widehat{\boldsymbol{V}}_{k,\perp}\widehat{\boldsymbol{V}}_{k,\perp}^\top\tilde{\boldsymbol{\beta}}^*\|_2 \le \|\boldsymbol{V}\boldsymbol{V}^\top - \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top\|_2\ \|\tilde{\boldsymbol{\beta}}^*\|_2. \tag{S12}$$

Bringing together (S3), (S9), *and* (S12). Collectively, we obtain

$$\begin{aligned}\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \le \|\boldsymbol{V}\boldsymbol{V}^\top - \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top\|_2^2\ \|\tilde{\boldsymbol{\beta}}^*\|_2^2 \\ + \frac{4}{\widehat{s}_k^2}\Big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2\ \|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\Big).\end{aligned} \tag{S13}$$

Key lemmas. We state the key lemmas bounding each of the terms on the right hand side of (S13). This will help us conclude the proof of Theorem 4.1. The proofs of these lemmas are presented in Sections C.1, C.2, C.3, C.4.

Lemma 2 *Consider the setup of Theorem 4.1, and PCR with parameter $k = r$. Then, for any $t > 0$, the following holds w.p. at least $1 - \exp(-t^2)$:*

$$\begin{aligned}\|\boldsymbol{U}\boldsymbol{U}^\top - \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{U}}_r^\top\|_2 \le C(K+1)(\gamma+1)\frac{\sqrt{n}+\sqrt{p}+t}{\rho s_r}, \\ \|\boldsymbol{V}\boldsymbol{V}^\top - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r^\top\|_2 \le C(K+1)(\gamma+1)\frac{\sqrt{n}+\sqrt{p}+t}{\rho s_r}.\end{aligned}$$

Here, $s_r > 0$ represents the r-th largest singular value of $\boldsymbol{X}$.

Lemma 3 *Consider PCR with parameter $k = r$ and $\rho \ge c(np)^{-1}\log^2(np)$. Then w.p. at least $1 - O(1/(np)^{10})$,*

$$\begin{aligned}&\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2 \\ &\quad\le C(K+1)^4(\gamma+1)^2\left(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r+\sqrt{r}\,\log(np)}{\rho^2}\right) + C\frac{\log(np)}{\rho\, p}.\end{aligned}$$

Lemma 4 *If $\rho \ge c(np)^{-1}\log^2(np)$, then for any k, we have w.p. at least $1 - O(1/(np)^{10})$,*

$$|\widehat{s}_k - s_k| \le C(K+1)(\gamma+1)\frac{\sqrt{n}+\sqrt{p}}{\rho} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}}s_k.$$

Lemma 5 *Given $\widetilde{\boldsymbol{Z}}^r$, the following holds w.p. at least $1 - O(1/(np)^{10})$ with respect to the randomness in $\boldsymbol{\varepsilon}$:*

$$\langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle \le \sigma^2 r + C\sigma\sqrt{\log(np)}\left(\sigma\sqrt{r} + \sigma\sqrt{\log(np)} + \|\tilde{\boldsymbol{\beta}}^*\|_1(\sqrt{n} + \|\widetilde{\boldsymbol{Z}}^r - \boldsymbol{X}\|_{2,\infty})\right).$$

Completing the proof of Theorem 4.1. Using Lemma 5, the following holds w.p. at least $1 - O(1/(np)^{10})$:

$$\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle$$

$$
\begin{aligned}
&\leq \|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2 \|\tilde{\boldsymbol{\beta}}^*\|_1^2 + C\sigma\sqrt{\log(np)}\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}\|\tilde{\boldsymbol{\beta}}^*\|_1 + C\sigma^2\log(np) \\
&\quad + C\sigma\sqrt{\log(np)}(\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1 + s\sigma\sqrt{r}) + \sigma^2 r \\
&\leq C\big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}\|\tilde{\boldsymbol{\beta}}^*\|_1 + \sigma\sqrt{\log(np)}\big)^2 + C\sigma\sqrt{\log(np)}(\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1 + \sigma\sqrt{r}) + \sigma^2 r \\
&\leq C\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + C\sigma^2(\log(np) + r) + C\sigma\sqrt{n\log(np)}\|\tilde{\boldsymbol{\beta}}^*\|_1. \qquad \text{(S14)}
\end{aligned}
$$

Using (S13) and (S14), we have w.p. at least $1 - O(1/(np)^{10})$,

$$
\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \leq \|\boldsymbol{V}\boldsymbol{V}^\top - \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top\|_2^2 \, \|\tilde{\boldsymbol{\beta}}^*\|_2^2 + C\frac{\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2}{\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \\
+ C\frac{\sigma^2(\log(np) + r)}{\widehat{s}_r^2} + C\frac{\sigma\sqrt{n\log(np)}}{\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1. \qquad \text{(S15)}
\end{aligned}
$$

Using Lemma 2 in (S15), we have w.p. at least $1 - O(1/(np)^{10})$,

$$
\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \leq C(K+1)^2(\gamma+1)^2\frac{n+p}{\rho^2 s_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_2^2 + C\frac{\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2}{\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \\
+ C\frac{\sigma^2(\log(np) + r)}{\widehat{s}_r^2} + C\frac{\sigma\sqrt{n\log(np)}}{\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1. \qquad \text{(S16)}
\end{aligned}
$$

Applying Lemma 4 with $k = r$ and recalling $\rho \geq c(np)^{-1}\log^2 np$ and $\mathsf{snr} \geq C(K+1)(\gamma+1)$,

$$
\begin{aligned}
\frac{|\widehat{s}_r - s_r|}{s_r} &\leq C(K+1)(\gamma+1)\frac{\sqrt{n} + \sqrt{p}}{\rho s_r} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} \\
&= \frac{C(K+1)(\gamma+1)}{\mathsf{snr}} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} \;\leq\; \frac{1}{2}.
\end{aligned}
$$

As a result,

$$
s_r/2 \leq \widehat{s}_r \leq 3s_r/2. \qquad \text{(S17)}
$$

Using the definition of snr as per (5) and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we have

$$
\frac{n+p}{\rho^2 s_r^2} \leq \frac{1}{\mathsf{snr}^2}. \qquad \text{(S18)}
$$

Using (S17), (S18) we obtain

$$
\begin{aligned}
\frac{\sigma^2(\log(np) + r)}{\widehat{s}_r^2} \;&\leq\; C\frac{\sigma^2\rho^2 r\log(np)}{\mathsf{snr}^2(n+p)} \leq C\frac{\sigma^2 r\log(np)}{\mathsf{snr}^2(n \vee p)}, \qquad \text{(S19)} \\
\frac{\sigma\sqrt{n\log(np)}}{\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1 \;&\leq C\frac{\sigma\rho^2\sqrt{n\log(np)}}{\mathsf{snr}^2(n+p)}\|\tilde{\boldsymbol{\beta}}^*\|_1 \leq C\frac{\sigma\sqrt{n\log(np)}}{\mathsf{snr}^2(n \vee p)}\|\tilde{\boldsymbol{\beta}}^*\|_1 \qquad \text{(S20)} \\
\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 &\leq C\frac{n\log(np)}{\mathsf{snr}^4(n+p)}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \leq C\frac{\log(np)}{\mathsf{snr}^4}\|\tilde{\boldsymbol{\beta}}^*\|_1^2, \qquad \text{(S21)} \\
\frac{r + \sqrt{r}\,\log(np)}{\rho^2\widehat{s}_r^2}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 &\leq C\frac{r\log(np)}{\mathsf{snr}^2(n+p)}\|\tilde{\boldsymbol{\beta}}^*\|_1^2, \qquad \text{(S22)}
\end{aligned}
$$

$$\frac{\log(np)}{\rho \widehat{s}_r^2 p}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \leq C\frac{\log(np)}{\mathsf{snr}^2 p(n+p)}\|\tilde{\boldsymbol{\beta}}^*\|_1^2. \tag{S23}$$

Plugging Lemma 3, (S19), (S20), (S21), (S22), (S23) into (S16), and simplifying completes the proof of (6) in Theorem 4.1.

It remains to establish (7). This result is first proved in Agarwal et al. (2023), cf. Lemma 19; we state a similar proof for completeness. By definition, $\tilde{\boldsymbol{\beta}}^* = \boldsymbol{X}^\dagger \boldsymbol{X}\boldsymbol{\beta}^*$. As such, it immediately follows that

$$\|\tilde{\boldsymbol{\beta}}^*\|_2 = \|\boldsymbol{X}^\dagger \boldsymbol{X}\boldsymbol{\beta}^*\|_2 \leq \|\boldsymbol{X}^\dagger\|_2 \cdot \|\boldsymbol{X}\boldsymbol{\beta}^*\|_2 \leq s_r^{-1} \cdot d\sqrt{n}.$$

The last inequality follows from our boundedness assumption on $\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle$ for all $i \leq n$.

Finally, the second part of (7) follows from the property $\|\boldsymbol{v}\|_1 \leq \sqrt{p}\|\boldsymbol{v}\|_2$ for any $\boldsymbol{v} \in \mathbb{R}^p$.

C.1 Proof of Lemma 2

Recall that $\boldsymbol{U}, \boldsymbol{V}$ denote the left and right singular vectors of $\boldsymbol{X}$ (equivalently, $\rho\boldsymbol{X}$), respectively; meanwhile, $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k$ denote the top k left and right singular vectors of $\widetilde{\boldsymbol{Z}}$ (equivalently, $\boldsymbol{Z}$), respectively. Further, observe that $\mathbb{E}[\boldsymbol{Z}] = \rho\boldsymbol{X}$ and let $\tilde{\boldsymbol{W}} = \boldsymbol{Z} - \rho\boldsymbol{X}$. To arrive at our result, we recall Wedin's Theorem (Wedin, 1972).

Theorem C.1 (Wedin's Theorem) *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times p}$, let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ and $\boldsymbol{B} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{V}}^\top$ be their respective SVDs. Let $\boldsymbol{U}_k, \boldsymbol{V}_k$ (respectively, $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{V}}_k$) correspond to the truncation of $\boldsymbol{U}, \boldsymbol{V}$ (respectively, $\widehat{\boldsymbol{U}}, \widehat{\boldsymbol{V}}$) that retains the columns corresponding to the top k singular values of $\boldsymbol{A}$ (respectively, $\boldsymbol{B}$). Let s_k denote the k-th singular value of $\boldsymbol{A}$. Then,*

$$\max\Big(\|\boldsymbol{U}_k\boldsymbol{U}_k^\top - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top\|_2, \|\boldsymbol{V}_k\boldsymbol{V}_k^\top - \widehat{\boldsymbol{V}}_k\widehat{\boldsymbol{V}}_k^\top\|_2\Big) \leq \frac{2\|\boldsymbol{A}-\boldsymbol{B}\|_2}{s_k - s_{k+1}}.$$

Using Theorem C.1 for $k = r$, it follows that

$$\max\Big(\|\boldsymbol{U}\boldsymbol{U}^\top - \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{U}}_r^\top\|_2, \|\boldsymbol{V}\boldsymbol{V}^\top - \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r^\top\|_2\Big) \leq \frac{2\|\tilde{\boldsymbol{W}}\|_2}{\rho s_r}, \tag{S24}$$

where s_r is the smallest nonzero singular value of $\boldsymbol{X}$. Next, we obtain a high probability bound on $\|\tilde{\boldsymbol{W}}\|_2$. To that end,

$$\frac{1}{n}\|\tilde{\boldsymbol{W}}\|_2^2 = \frac{1}{n}\|\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}\|_2 \leq \frac{1}{n}\|\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}} - \mathbb{E}[\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}]\|_2 + \frac{1}{n}\|\mathbb{E}[\tilde{\boldsymbol{W}}^\top\tilde{\boldsymbol{W}}]\|_2. \tag{S25}$$

We bound the two terms in (S25) separately. We recall the following lemma, which is a direct extension of Theorem 4.6.1 of Vershynin (2018) for the non-isotropic setting, and we present its proof for completeness in Section C.5.

Lemma 6 (Independent sub-gaussian rows) *Let $\boldsymbol{A}$ be an $n \times p$ matrix whose rows A_i are independent, mean zero, sub-gaussian random vectors in $\mathbb{R}^p$ with second moment matrix $\boldsymbol{\Sigma} = (1/n)\mathbb{E}[\boldsymbol{A}^\top\boldsymbol{A}]$. Then for any $t \geq 0$, the following holds w.p. at least $1 - \exp(-t^2)$:*

$$\|\frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{\Sigma}\|_2 \leq K^2\max(\delta, \delta^2), \quad \text{where } \delta = C\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}; \tag{S26}$$

here, $K = \max_i \|A_i\|_{\psi_2}$.

The matrix $\tilde{\boldsymbol{W}} = \boldsymbol{Z} - \rho \boldsymbol{X}$ has independent rows by Assumption 3.2. We state the following Lemma about the distribution property of the rows of $\tilde{\boldsymbol{W}}$, the proof of which can be found in Section C.6.

Lemma 7 *Let Assumption 3.2 hold. Then, $\boldsymbol{z}_i - \rho \boldsymbol{x}_i$ is a sequence of independent, mean zero, sub-gaussian random vectors satisfying $\|\boldsymbol{z}_i - \rho \boldsymbol{x}_i\|_{\psi_2} \le C(K+1)$.*

From Lemmas 6 and 7, w.p. at least $1 - \exp\bigl(-t^2\bigr)$,

$$\frac{1}{n}\|\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}} - \mathbb{E}[\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}]\|_2 \le C(K+1)^2 \left(1 + \frac{p}{n} + \frac{t^2}{n}\right). \tag{S27}$$

Finally, we claim the following bound on $\|\mathbb{E}[\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}]\|_2$, the proof of which is in Section C.7.

Lemma 8 *Let Assumption 3.2 hold. Then, we have*

$$\|\mathbb{E}[\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}]\|_2 \le C(K+1)^2 n(\rho - \rho^2) + n\rho^2\gamma^2.$$

From (S25), (S27) and Lemma 8, we have w.p. at least $1 - \exp\bigl(-t^2\bigr)$ for any $t > 0$

$$\|\tilde{\boldsymbol{W}}\|_2^2 \le C(K+1)^2(n + p + t^2) + n(\rho(1-\rho)(K+1)^2 + \rho^2\gamma^2).$$

For this, we conclude the following lemma.

Lemma 9 *For any $t > 0$, the following holds w.p. at least $1 - \exp\bigl(-t^2\bigr)$:*

$$\|\boldsymbol{Z} - \rho \boldsymbol{X}\|_2 \le C(K+1)(\gamma+1)(\sqrt{n} + \sqrt{p} + t).$$

Using the above and (S24), we conclude the proof of Lemma 2.

C.2 Proof of Lemma 3

We want to bound $\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2$. To that end, let $\Delta_j = \boldsymbol{X}_{\cdot j} - \widetilde{\boldsymbol{Z}}^k_{\cdot j}$ for any $j \in [p]$. Our interest is in bounding $\|\Delta_j\|_2^2$ for all $j \in [p]$. Consider,

$$\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \boldsymbol{X}_{\cdot j} = (\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}) + (\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}).$$

Now, note that $\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}$ belongs to the subspace spanned by column vectors of $\widehat{\boldsymbol{U}}_k$, while $\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}$ belongs to its orthogonal complement with respect to $\mathbb{R}^n$. As a result,

$$\|\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \boldsymbol{X}_{\cdot j}\|_2^2 = \|\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}\|_2^2 + \|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}\|_2^2. \tag{S28}$$

Bounding $\|\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}\|_2^2$. Recall that $\widetilde{\boldsymbol{Z}} = (1/\widehat{\rho})\boldsymbol{Z} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{V}}^\top$, and hence $\boldsymbol{Z} = \widehat{\rho}\widehat{\boldsymbol{U}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{V}}^\top$. Consequently,

$$\frac{1}{\widehat{\rho}}\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{Z}_{\cdot j} = \frac{1}{\widehat{\rho}}\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{Z}\boldsymbol{e}_j = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \widehat{\boldsymbol{U}}\widehat{\boldsymbol{S}}\widehat{\boldsymbol{V}}^\top \boldsymbol{e}_j$$

$$= \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top \boldsymbol{e}_j = \widetilde{\boldsymbol{Z}}_{\cdot j}^k.$$

Therefore, we have

$$\begin{aligned}\widetilde{\boldsymbol{Z}}_{\cdot j}^k - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} &= \frac{1}{\widehat{\rho}} \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{Z}_{\cdot j} - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} \\ &= \frac{1}{\widehat{\rho}} \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}) + \Big(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\Big) \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}_{\cdot j}^k - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}\|_2^2 &\le \frac{2}{\widehat{\rho}^2} \|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 + 2\Big(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\Big)^2 \|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}\|_2^2 \\ &\le \frac{2}{\widehat{\rho}^2} \|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 + 2\Big(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\Big)^2 \|\boldsymbol{X}_{\cdot j}\|_2^2,\end{aligned}$$

where we have used the fact that $\|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top\|_2 = 1$. Recall that $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ represents the left singular vectors of $\boldsymbol{X}$. Thus,

$$\begin{aligned}\|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 &\le 2\|(\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top - \boldsymbol{U}\boldsymbol{U}^\top)(\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 + 2\|\boldsymbol{U}\boldsymbol{U}^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 \\ &\le 2\|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_2^2 \, \|\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}\|_2^2 + 2\|\boldsymbol{U}\boldsymbol{U}^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2.\end{aligned}$$

By Assumption 3.3, we have that $\|\boldsymbol{X}_{\cdot j}\|_2^2 \le n$. This yields

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}_{\cdot j}^k - \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}\|_2^2 \le \frac{4}{\widehat{\rho}^2} &\|\widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_2^2 \, \|\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}\|_2^2 \\ &+ \frac{4}{\widehat{\rho}^2} \|\boldsymbol{U}\boldsymbol{U}^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 + 2n\Big(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\Big)^2. \qquad \text{(S29)}\end{aligned}$$

We now state Lemmas 10 and 11. Their proofs are in Sections C.8 and C.9, respectively.

Lemma 10 *For any $\alpha > 1$,*

$$\mathbb{P}\left(\rho/\alpha \le \widehat{\rho} \le \alpha\rho\right) \ge 1 - 2\exp\Big(-\frac{(\alpha-1)^2 np\rho}{2\alpha^2}\Big).$$

Therefore, for $\rho \ge c\frac{\log^2 np}{np}$, we have w.p. $1 - O(1/(np)^{10})$

$$\frac{\rho}{2} \le \widehat{\rho} \le 2\rho \quad \textit{and} \quad \Big(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\Big)^2 \le C\frac{\log(np)}{\rho np}.$$

Lemma 11 *Consider any matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times \ell}$ with $1 \le \ell \le n$ such that its columns $\boldsymbol{Q}_{\cdot j}$ for $j \in [\ell]$ are orthonormal vectors. Then for any $t > 0$,*

$$\begin{aligned}&\mathbb{P}\Big(\max_{j \in [p]} \Big\|\boldsymbol{Q}\boldsymbol{Q}^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\Big\|_2^2 \ge \ell C(K+1)^2 + t\Big) \\ &\qquad \le p \cdot \exp\Big(-c\min\Big(\frac{t^2}{C(K+1)^4 \ell}, \frac{t}{C(K+1)^2}\Big)\Big).\end{aligned}$$

Subsequently, w.p. $1 - O(1/(np)^{10})$,

$$\max_{j \in [p]} \Big\|\boldsymbol{Q}\boldsymbol{Q}^\top (\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\Big\|_2^2 \le C(K+1)^2(\ell + \sqrt{\ell}\log(np)).$$

Both terms $\|\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}\|_2^2$ and $\|\boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2$ can be bounded by Lemma 11: for the first term $\boldsymbol{Q} = \mathbf{I}$, and for the second term $\boldsymbol{Q} = \boldsymbol{U}$. In summary, w.p. $1 - O(1/(np)^{10})$, we have

$$\max_{j\in[p]} \|\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j}\|_2^2 \le C(K+1)^2(n + \sqrt{n}\log(np)), \tag{S30}$$

and

$$\max_{j\in[p]} \|\boldsymbol{U}\boldsymbol{U}^\top(\boldsymbol{Z}_{\cdot j} - \rho \boldsymbol{X}_{\cdot j})\|_2^2 \le C(K+1)^2(r + \sqrt{r}\log(np)). \tag{S31}$$

Using (S29), (S30), (S31), and Lemmas 2 and 10 with $k = r$, we conclude that w.p. $1 - O(1/(np)^{10})$,

$$\begin{aligned}&\max_{j\in[p]} \|\widetilde{\boldsymbol{Z}}^k_{\cdot j} - \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j}\|_2^2 \\ &\le C(K+1)^4(\gamma+1)^2\left(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r+\sqrt{r}\,\log(np)}{\rho^2}\right) + C\frac{\log(np)}{\rho\, p}\end{aligned} \tag{S32}$$

Bounding $\|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}\|_2^2$. Recalling $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, we obtain $\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X}_{\cdot j} = \boldsymbol{X}_{\cdot j}$ since $\boldsymbol{U}\boldsymbol{U}^\top$ is the projection onto the column space of $\boldsymbol{X}$. Therefore,

$$\begin{aligned}\|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}\|_2^2 &= \|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{X}_{\cdot j}\|_2^2 \\ &\le \|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top - \boldsymbol{U}\boldsymbol{U}^\top\|_2^2\, \|\boldsymbol{X}_{\cdot j}\|_2^2.\end{aligned}$$

Using Property 3.3, note that $\|\boldsymbol{X}_{\cdot j}\|_2^2 \le n$. Thus using Lemma 2 with $k = r$, we have that w.p. at least $1 - O(1/(np)^{10})$, we have

$$\|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}_{\cdot j} - \boldsymbol{X}_{\cdot j}\|_2^2 \le C\frac{n(n+p)}{\rho^2 s_r^2}. \tag{S33}$$

Concluding. From (S28), (S32), and (S33), we claim w.p. at least $1 - O(1/(np)^{10})$

$$\begin{aligned}&\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^k\|_{2,\infty}^2 \\ &\quad\le C(K+1)^4(\gamma+1)^2\left(\frac{(n+p)(n+\sqrt{n}\,\log(np))}{\rho^4 s_r^2} + \frac{r+\sqrt{r}\,\log(np)}{\rho^2}\right) + C\frac{\log(np)}{\rho\, p}.\end{aligned}$$

This completes the proof of Lemma 3.

C.3 Proof of Lemma 4

To bound $\widehat{s}_k$, we recall Weyl's inequality.

Lemma 12 (Weyl's inequality) *Given $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, let σ_i and $\widehat{\sigma}_i$ be the i-th singular values of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, in decreasing order and repeated by multiplicities. Then for all $i \in [m \wedge n]$,*

$$|\sigma_i - \widehat{\sigma}_i| \le \|\boldsymbol{A} - \boldsymbol{B}\|_2.$$

Let $\tilde{s}_k$ be the k-th singular value of $\boldsymbol{Z}$. Then, $\widehat{s}_k = (1/\widehat{\rho})\tilde{s}_k$ since it is the k-th singular value of $\widetilde{\boldsymbol{Z}} = (1/\widehat{\rho})\boldsymbol{Z}$. By Lemma 12, we have

$$|\tilde{s}_k - \rho s_k| \leq \|\boldsymbol{Z} - \rho \boldsymbol{X}\|_2;$$

recall that s_k is the k-th singular value of $\boldsymbol{X}$. As a result,

$$\begin{aligned}
|\widehat{s}_k - s_k| &= \frac{1}{\widehat{\rho}}|\tilde{s}_k - \widehat{\rho} s_k| \\
&\leq \frac{1}{\widehat{\rho}}|\tilde{s}_k - \rho s_k| + \frac{|\rho - \widehat{\rho}|}{\widehat{\rho}} s_k \\
&\leq \frac{\|\boldsymbol{Z} - \rho \boldsymbol{X}\|_2}{\widehat{\rho}} + \frac{|\rho - \widehat{\rho}|}{\widehat{\rho}} s_k.
\end{aligned}$$

From Lemma 9 and Lemma 10, it follows that w.p. at least $1 - O(1/(np)^{10})$,

$$|\widehat{s}_k - s_k| \leq C(K+1)(\gamma+1)\frac{\sqrt{n}+\sqrt{p}}{\rho} + C\frac{\sqrt{\log(np)}}{\sqrt{\rho\, np}} s_k.$$

This completes the proof of Lemma 4.

C.4 Proof of Lemma 5

We need to bound $\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle$. To that end, we recall that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^\top y$, $\widetilde{\boldsymbol{Z}}^k = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top$, and $\boldsymbol{y} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}^* + \boldsymbol{\varepsilon}$. Thus,

$$\widetilde{\boldsymbol{Z}}^k \widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top \widehat{\boldsymbol{V}}_k \widehat{\boldsymbol{S}}_k^{-1} \widehat{\boldsymbol{U}}_k^\top \boldsymbol{y} = \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}^* + \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{\varepsilon}.$$

Therefore,

$$\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle = \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle + \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle - \langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{S}}_k \widehat{\boldsymbol{V}}_k^\top \tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle. \tag{S34}$$

Now, $\boldsymbol{\varepsilon}$ is independent of $\widehat{\boldsymbol{U}}_k, \widehat{\boldsymbol{S}}_k, \widehat{\boldsymbol{V}}_k$ since $\widetilde{\boldsymbol{Z}}^k$ is determined by $\boldsymbol{Z}$, which is independent of $\boldsymbol{\varepsilon}$. As a result,

$$\begin{aligned}
\mathbb{E}\big[\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle\big] &= \mathbb{E}\big[\boldsymbol{\varepsilon}^\top \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{\varepsilon}\big] \\
&= \mathbb{E}\big[\operatorname{tr}\Big(\boldsymbol{\varepsilon}^\top \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{\varepsilon}\Big)\big] = \mathbb{E}\big[\operatorname{tr}\Big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top\Big)\big] \\
&= \operatorname{tr}\Big(\mathbb{E}\big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\big] \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top\Big) \leq C \operatorname{tr}\Big(\sigma^2 \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top\Big) \\
&= C\sigma^2 \|\widehat{\boldsymbol{U}}_k\|_F^2 = C\sigma^2 k.
\end{aligned} \tag{S35}$$

Therefore, it follows that

$$\mathbb{E}[\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle] \leq C\sigma^2 k, \tag{S36}$$

where we used the fact $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$. To obtain a high probability bound, using Lemma 16 it follows that for any $t > 0$

$$\mathbb{P}\Big(\langle \widehat{\boldsymbol{U}}_k \widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle \geq t\Big) \leq \exp\Big(-\frac{ct^2}{n\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \sigma^2}\Big) \tag{S37}$$

due to Assumption 3.1, and

$$\|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2 \le \|\boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2 \le \|\boldsymbol{X}\|_{2,\infty}\|\tilde{\boldsymbol{\beta}}^*\|_1 \le \sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1;$$

note that we have used the fact that $\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top$ is a projection matrix and $\|\boldsymbol{X}\|_{2,\infty} \le \sqrt{n}$ due to Assumption 3.3. Similarly, for any $t > 0$

$$\mathbb{P}\Big(\langle \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{S}}_k\widehat{\boldsymbol{V}}_k^\top\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle \ge t\Big) \le \exp\Big(-\frac{ct^2}{\sigma^2(n + \|\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}\|_{2,\infty}^2)\|\tilde{\boldsymbol{\beta}}^*\|_1^2}\Big), \tag{S38}$$

due to Assumption 3.1, and

$$\begin{aligned}\|\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{S}}_k\widehat{\boldsymbol{V}}_k^\top\tilde{\boldsymbol{\beta}}^*\|_2 = \|(\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\tilde{\boldsymbol{\beta}}^* + \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2 \ &\le \|(\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X})\tilde{\boldsymbol{\beta}}^*\|_2 + \|\boldsymbol{X}\tilde{\boldsymbol{\beta}}^*\|_2 \\ &\le \big(\|\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}\|_{2,\infty} + \|\boldsymbol{X}\|_{2,\infty}\big)\|\tilde{\boldsymbol{\beta}}^*\|_1.\end{aligned}$$

Finally, using Lemma 17 and (S36), it follows that for any $t > 0$

$$\mathbb{P}\Big(\langle \widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle \ge \sigma^2 k + t\Big) \le \exp\Big(-c\min\Big(\frac{t^2}{k\sigma^4}, \frac{t}{\sigma^2}\Big)\Big), \tag{S39}$$

since $\widehat{\boldsymbol{U}}_k\widehat{\boldsymbol{U}}_k^\top$ is a projection matrix and by Assumption 3.1.

From (S34), (S37), (S38), and (S39), we conclude that w.p. at least $1 - O(1/(np)^{10})$,

$$\langle \widetilde{\boldsymbol{Z}}^k(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle \le \sigma^2 k + C\sigma\sqrt{\log(np)}\Big(\sigma\sqrt{k} + \sigma\sqrt{\log(np)} + \|\tilde{\boldsymbol{\beta}}^*\|_1(\sqrt{n} + \|\widetilde{\boldsymbol{Z}}^k - \boldsymbol{X}\|_{2,\infty})\Big).$$

This completes the proof of Lemma 5.

C.5 Proof of Lemma 6

As mentioned earlier, the proof presented here is a natural extension of that for Theorem 4.6.1 in Vershynin (2018) for the non-isotropic setting. Recall that

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\in S^{p-1}, \boldsymbol{y}\in S^{n-1}} \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle,$$

where S^{p-1}, S^{n-1} denote the unit spheres in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively. We start by bounding the quadratic term $\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle$ for a finite set $\boldsymbol{x}, \boldsymbol{y}$ obtained by placing 1/4-net on the unit spheres, and then use the bound on them to bound $\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}\rangle$ for all $\boldsymbol{x}, \boldsymbol{y}$ over the spheres.

Step 1: Approximation. We will use Corollary 4.2.13 of Vershynin (2018) to establish a 1/4-net of $\mathcal{N}$ of the unit sphere S^{p-1} with cardinality $|\mathcal{N}| \le 9^p$. Applying Lemma 4.4.1 of Vershynin (2018), we obtain

$$\|\frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{\Sigma}\|_2 \le 2\max_{\boldsymbol{x}\in\mathcal{N}}\Big|\langle(\frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{\Sigma})\boldsymbol{x}, \boldsymbol{x}\rangle\Big| = 2\max_{\boldsymbol{x}\in\mathcal{N}}\Big|\frac{1}{n}\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}\Big|.$$

To achieve our desired result, it remains to show that

$$\max_{\boldsymbol{x}\in\mathcal{N}}\Big|\frac{1}{n}\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}\Big| \le \frac{\epsilon}{2},$$

where $\epsilon = K^2 \max(\delta, \delta^2)$.

Step 2: Concentration. Let us fix a unit vector $\boldsymbol{x} \in S^{p-1}$ and write

$$\|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} = \sum_{i=1}^n \left(\langle \boldsymbol{A}_{i,\cdot}, \boldsymbol{x} \rangle^2 - \mathbb{E}[\langle \boldsymbol{A}_{i,\cdot}, \boldsymbol{x} \rangle^2] \right) =: \sum_{i=1}^n \left(Y_i^2 - \mathbb{E}[Y_i^2] \right).$$

Since the rows of $\boldsymbol{A}$ are assumed to be independent sub-gaussian random vectors with $\|\boldsymbol{A}_{i,\cdot}\|_{\psi_2} \leq K$, it follows that $Y_i = \langle \boldsymbol{A}_{i,\cdot}, \boldsymbol{x} \rangle$ are independent sub-gaussian random variables with $\|Y_i\|_{\psi_2} \leq K$. Therefore, $Y_i^2 - \mathbb{E}[Y_i^2]$ are independent, mean zero, sub-exponential random variables with

$$\|Y_i^2 - \mathbb{E}[Y_i^2]\|_{\psi_1} \leq C \|Y_i^2\|_{\psi_1} \leq C \|Y_i\|_{\psi_2}^2 \leq C K^2.$$

As a result, we can apply Bernstein's inequality (see Theorem E.1) to obtain

$$\begin{aligned}
\mathbb{P}\Big(\Big| \frac{1}{n} \|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \Big| \geq \frac{\epsilon}{2} \Big) &= \mathbb{P}\Big(\Big| \frac{1}{n} \sum_{i=1}^n (Y_i^2 - \mathbb{E}[Y_i^2]) \Big| \geq \frac{\epsilon}{2} \Big) \\
&\leq 2 \exp\left(-c \min\left(\frac{\epsilon^2}{K^4}, \frac{\epsilon}{K^2} \right) n \right) \\
&= 2 \exp\big(-c \delta^2 n \big) \\
&\leq 2 \exp\big(-c C^2 (p + t^2) \big),
\end{aligned}$$

where the last inequality follows from the definition of δ in (S26) and because $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$.

Step 3: Union bound. We now apply a union bound over all elements in the net,

$$\mathbb{P}\Big(\max_{\boldsymbol{x} \in \mathcal{N}} \Big| \frac{1}{n} \|\boldsymbol{A}\boldsymbol{x}\|_2^2 - \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \Big| \geq \frac{\epsilon}{2} \Big) \leq 9^p \cdot 2 \exp\big(-c C^2 (p + t^2) \big) \leq 2 \exp\big(-t^2 \big),$$

for large enough C. This concludes the proof.

C.6 Proof of Lemma 7

Recall that $\boldsymbol{z}_i = (\boldsymbol{x}_i + \boldsymbol{w}_i) \circ \boldsymbol{\pi}_i$, where $\boldsymbol{w}_i$ is an independent mean zero subgaussian vector with $\|\boldsymbol{w}_i\|_{\psi_2} \leq K$ and $\boldsymbol{\pi}_i$ is a vector of independent Bernoulli variables with parameter ρ. Hence, $\mathbb{E}[\boldsymbol{z}_i - \rho \boldsymbol{x}_i] = \boldsymbol{0}$ and is independent across $i \in [n]$. The only remaining item is a bound on $\|\boldsymbol{z}_i - \rho \boldsymbol{x}_i\|_{\psi_2}$. To that end, note that

$$\begin{aligned}
\|\boldsymbol{z}_i - \rho \boldsymbol{x}_i\|_{\psi_2} &= \|\boldsymbol{x}_i \circ \boldsymbol{\pi}_i + \boldsymbol{w}_i \circ \boldsymbol{\pi}_i - \rho \boldsymbol{x}_i\|_{\psi_2} \\
&\leq \|\boldsymbol{x}_i \circ (\rho \boldsymbol{1} - \boldsymbol{\pi}_i)\|_{\psi_2} + \|\boldsymbol{w}_i \circ \boldsymbol{\pi}_i\|_{\psi_2}.
\end{aligned}$$

Now, $(\rho \boldsymbol{1} - \boldsymbol{\pi}_i)$ is independent, zero mean random vector whose absolute value is bounded by 1, and is component-wise multiplied by $\boldsymbol{x}_i$ which are bounded in absolute value by 1 as per Assumption 3.3. That is, $\boldsymbol{x}_i \circ (\rho \boldsymbol{1} - \boldsymbol{\pi}_i)$ is a zero mean random vector where each component is independent and bounded in absolute value by 1. That is, $\|\cdot\|_{\psi_2} \leq C$.

For $\boldsymbol{w}_i \circ \boldsymbol{\pi}_i$, note that $\boldsymbol{w}_i$ and $\boldsymbol{\pi}_i$ are independent vectors and the coordinates of $\boldsymbol{\pi}_i$ have support $\{0, 1\}$. Therefore, from Lemma 13, it follows that $\|\boldsymbol{w}_i \circ \boldsymbol{\pi}_i\|_{\psi_2} \leq \|\boldsymbol{w}_i\|_{\psi_2} \leq K$ by Assumption 3.2. The proof of Lemma 7 is complete by choosing a large enough C.

Lemma 13 *Suppose that $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{P} \in \{0,1\}^n$ are independent random vectors. Then,*

$$\|\boldsymbol{Y} \circ \boldsymbol{P}\|_{\psi_2} \leq \|\boldsymbol{Y}\|_{\psi_2}.$$

Proof Given a binary vector $\boldsymbol{P} \in \{0,1\}^n$, let $I_{\boldsymbol{P}} = \{i \in [n] : P_i = 1\}$. Observe that

$$\boldsymbol{Y} \circ \boldsymbol{P} = \sum_{i \in I_{\boldsymbol{P}}} \boldsymbol{e}_i \otimes \boldsymbol{e}_i \boldsymbol{Y}.$$

Here, $\circ$ denotes the Hadamard product (entry-wise product) of two matrices. By definition of the ψ_2-norm,

$$\|\boldsymbol{Y}\|_{\psi_2} = \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \|\boldsymbol{u}^\top \boldsymbol{Y}\|_{\psi_2} = \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{\boldsymbol{Y}}[\exp\Big(|\boldsymbol{u}^\top \boldsymbol{Y}|^2/t^2\Big)] \leq 2\}.$$

Let $\boldsymbol{u}_0 \in \mathbb{S}^{n-1}$ denote the maximum-achieving unit vector (such a $\boldsymbol{u}_0$ exists because $\inf\{\cdots\}$ is continuous with respect to $\boldsymbol{u}$ and $\mathbb{S}^{n-1}$ is compact). Now,

$$\begin{aligned}
\|\boldsymbol{Y} \circ \boldsymbol{P}\|_{\psi_2} &= \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \|\boldsymbol{u}^\top \boldsymbol{Y} \circ \boldsymbol{P}\|_{\psi_2} \\
&= \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{\boldsymbol{Y},\boldsymbol{P}}[\exp\Big(|\boldsymbol{u}^\top \boldsymbol{Y} \circ \boldsymbol{P}|^2/t^2\Big)] \leq 2\} \\
&= \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{\boldsymbol{P}}[\mathbb{E}_{\boldsymbol{Y}}[\exp\Big(|\boldsymbol{u}^\top \boldsymbol{Y} \circ \boldsymbol{P}|^2/t^2\Big) \mid \boldsymbol{P}]] \leq 2\} \\
&= \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{\boldsymbol{P}}[\mathbb{E}_{\boldsymbol{Y}}[\exp(|\boldsymbol{u}^\top \sum_{i \in I_{\boldsymbol{P}}} \boldsymbol{e}_i \otimes \boldsymbol{e}_i \boldsymbol{Y}|^2/t^2) \mid \boldsymbol{P}]] \leq 2\} \\
&= \sup_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \inf\{t > 0 : \mathbb{E}_{\boldsymbol{P}}[\mathbb{E}_{\boldsymbol{Y}}[\exp(|(\sum_{i \in I_{\boldsymbol{P}}} \boldsymbol{e}_i \otimes \boldsymbol{e}_i \boldsymbol{u})^\top \boldsymbol{Y}|^2/t^2) \mid \boldsymbol{P}]] \leq 2\}.
\end{aligned}$$

For any $\boldsymbol{u} \in \mathbb{S}^{n-1}$, observe that

$$\mathbb{E}_{\boldsymbol{Y}}[\exp(|(\sum_{i \in I_P} \boldsymbol{e}_i \otimes \boldsymbol{e}_i \boldsymbol{u})^\top \boldsymbol{Y}|^2/t^2) \mid \boldsymbol{P}] \leq \mathbb{E}_{\boldsymbol{Y}}[\exp\Big(|\boldsymbol{u}_0^\top \boldsymbol{Y}|^2/t^2\Big)].$$

Therefore, taking supremum over $\boldsymbol{u} \in \mathbb{S}^{n-1}$, we obtain

$$\|\boldsymbol{Y} \circ \boldsymbol{P}\|_{\psi_2} \leq \|\boldsymbol{Y}\|_{\psi_2}.$$

■

C.7 Proof of Lemma 8

Consider

$$\mathbb{E}[\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}] = \sum_{i=1}^{n} \mathbb{E}[(\boldsymbol{z}_i - \rho \boldsymbol{x}_i) \otimes (\boldsymbol{z}_i - \rho \boldsymbol{x}_i)]$$

$$= \sum_{i=1}^{n} \mathbb{E}[\boldsymbol{z}_i \otimes \boldsymbol{z}_i] - \rho^2(\boldsymbol{x}_i \otimes \boldsymbol{x}_i)$$
$$= \sum_{i=1}^{n} (\rho - \rho^2)\mathrm{diag}(\boldsymbol{x}_i \otimes \boldsymbol{x}_i) + (\rho - \rho^2)\mathrm{diag}(\mathbb{E}[\boldsymbol{w}_i \otimes \boldsymbol{w}_i]) + \rho^2\mathbb{E}[\boldsymbol{w}_i \otimes \boldsymbol{w}_i].$$

Note that $\|\mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X})\|_2 \leq n$ due to Assumption 3.3. Using Assumption 3.2, it follows that $\|\mathrm{diag}(\mathbb{E}[\boldsymbol{w}_i \otimes \boldsymbol{w}_i])\|_2 \leq CK^2$. By Assumption 3.2, we have $\|\mathbb{E}[\boldsymbol{w}_i \otimes \boldsymbol{w}_i]\|_2 \leq \gamma^2$. Therefore,

$$\|\mathbb{E}[\tilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{W}}]\|_2 \leq Cn(\rho - \rho^2)(K+1)^2 + n\rho^2\gamma^2.$$

This completes the proof of Lemma 8.

C.8 Proof of Lemma 10

By the Binomial Chernoff bound, for $\alpha > 1$,

$$\mathbb{P}\left(\widehat{\rho} > \alpha\rho\right) \leq \exp\left(-\frac{(\alpha-1)^2}{\alpha+1} np\rho\right) \quad \text{and} \quad \mathbb{P}\left(\widehat{\rho} < \rho/\alpha\right) \leq \exp\left(-\frac{(\alpha-1)^2}{2\alpha^2} np\rho\right).$$

By the union bound,

$$\mathbb{P}\left(\rho/\alpha \leq \widehat{\rho} \leq \alpha\rho\right) \geq 1 - \mathbb{P}\left(\widehat{\rho} > \alpha\rho\right) - \mathbb{P}\left(\widehat{\rho} < \rho/\alpha\right).$$

Noticing $\alpha + 1 < 2\alpha < 2\alpha^2$ for all $\alpha > 1$, we obtain the desired bound claimed in Lemma 10. To complete the remaining claim of Lemma 10, we consider an α that satisfies

$$(\alpha - 1)^2 \leq C\frac{\log(np)}{\rho np},$$

for a constant $C > 0$. Thus,

$$1 - C\frac{\sqrt{\log(np)}}{\sqrt{\rho np}} \leq \alpha \leq 1 + C\frac{\sqrt{\log(np)}}{\sqrt{\rho np}}.$$

Then, with $\rho \geq c\frac{\log^2 np}{np}$, we have that $\alpha \leq 2$. Further by choosing $C > 0$ large enough, we have

$$\frac{(\rho - \widehat{\rho})^2}{\widehat{\rho}^2} \leq C\frac{\log(np)}{\rho np}.$$

holds w.p. at least $1 - O(1/(np)^{10})$. This completes the proof of Lemma 10.

C.9 Proof of Lemma 11

By definition $\boldsymbol{Q}\boldsymbol{Q}^\top \in \mathbb{R}^{n\times n}$ is a rank ℓ matrix. Since $\boldsymbol{Q}$ has orthonormal column vectors, the projection operator has $\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_2 = 1$ and $\|\boldsymbol{Q}\boldsymbol{Q}^\top\|_F^2 = \ell$. For a given $j \in [p]$, the random vector $\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}$ is such that it has zero mean, independent components that are sub-gaussian by Assumption 3.2. For any $i \in [n], j \in [p]$, we have by property of ψ_2 norm, $\|z_{ij} - \rho x_{ij}\|_{\psi_2} \leq \|\boldsymbol{z}_i - \rho\boldsymbol{x}_i\|_{\psi_2}$ which is bounded by $C(K+1)$ using Lemma 7. Recall the Hanson-Wright inequality (Vershynin (2018)):

Theorem C.2 (Hanson-Wright inequality) *Let $\boldsymbol{\zeta} \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates. Let $\boldsymbol{A}$ be an $n \times n$ matrix. Then for any $t > 0$,*

$$\mathbb{P}\left(\left|\boldsymbol{\zeta}^\top \boldsymbol{A}\boldsymbol{\zeta} - \mathbb{E}[\boldsymbol{\zeta}^\top \boldsymbol{A}\boldsymbol{\zeta}]\right| \geq t\right) \leq 2\exp\Big(-c\min\Big(\frac{t^2}{L^4\|\boldsymbol{A}\|_F^2}, \frac{t}{L^2\|\boldsymbol{A}\|_2}\Big)\Big),$$

where $L = \max_{i\in[n]} \|\zeta_i\|_{\psi_2}$.

Now with $\boldsymbol{\zeta} = \boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}$ and the fact that $\boldsymbol{Q}^\top\boldsymbol{Q} = \mathbf{I} \in \mathbb{R}^{\ell\times\ell}$, $\|\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\zeta}\|_2^2 = \boldsymbol{\zeta}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\zeta}$. Therefore, by Theorem C.2, for any $t > 0$,

$$\|\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\zeta}\|_2^2 \leq \mathbb{E}\big[\boldsymbol{\zeta}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\zeta}\big] + t,$$

w.p. at least $1 - \exp\Big(-c\min\big(\frac{t}{C(K+1)^2}, \frac{t^2}{C(K+1)^4\ell}\big)\Big)$. Now,

$$\begin{aligned}
\mathbb{E}\big[\boldsymbol{\zeta}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\zeta}\big] &= \sum_{m=1}^{\ell} \mathbb{E}\big[(\boldsymbol{Q}_{\cdot m}^\top\boldsymbol{\zeta})^2\big] \\
&\overset{(a)}{=} \sum_{m=1}^{\ell} \mathbb{V}\mathrm{ar}(\boldsymbol{Q}_{\cdot m}^\top\boldsymbol{\zeta}) \\
&\overset{(b)}{=} \sum_{m=1}^{\ell}\sum_{i=1}^{n} \boldsymbol{Q}_{im}^2 \mathbb{V}\mathrm{ar}(\zeta_i) \\
&\overset{(c)}{\leq} C(K+1)^2\ell,
\end{aligned}$$

where $\boldsymbol{\zeta} = \boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}$, and hence (a) follows from $\mathbb{E}[\boldsymbol{\zeta}] = \mathbb{E}[\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j}] = \mathbf{0}$, (b) follows from $\boldsymbol{\zeta}$ having independent components and (c) follows from each component of $\boldsymbol{\zeta}$ having ψ_2-norm bounded by $C(K+1)$. Therefore, it follows by union bound that for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\Big(\max_{j\in[p]} \Big\|\boldsymbol{Q}\boldsymbol{Q}^\top(\boldsymbol{Z}_{\cdot j} - \rho\boldsymbol{X}_{\cdot j})\Big\|_2^2 \geq \ell C(K+1)^2 + t\Big) \\
&\quad\leq p\cdot\exp\Big(-c\min\Big(\frac{t^2}{C(K+1)^4\ell}, \frac{t}{C(K+1)^2}\Big)\Big).
\end{aligned}$$

This completes the proof of Lemma 11.

Appendix D. Proof of Theorem 4.2

Recall that $\boldsymbol{X}'$ and $\boldsymbol{Z}'$ denote the latent and observed testing covariates, respectively. We denote the SVD of the former as $\boldsymbol{X}' = \boldsymbol{U}'\boldsymbol{S}'\boldsymbol{V}'^\top$. Let s'_ℓ be the ℓ-th singular value of $\boldsymbol{X}'$. Further, recall that $\widetilde{\boldsymbol{Z}}' = (1/\widehat{\rho}')\boldsymbol{Z}'$, and its rank ℓ truncation is denoted as $\widetilde{\boldsymbol{Z}}'^\ell$. Our interest is in bounding $\|\widetilde{\boldsymbol{Z}}'^\ell\widehat{\boldsymbol{\beta}} - \boldsymbol{X}'\tilde{\boldsymbol{\beta}}^*\|_2$. Towards this, consider

$$\begin{aligned}
\|\widetilde{\boldsymbol{Z}}'^\ell\widehat{\boldsymbol{\beta}} - \boldsymbol{X}'\tilde{\boldsymbol{\beta}}^*\|_2^2 &= \|\widetilde{\boldsymbol{Z}}'^\ell\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{Z}}'^\ell\tilde{\boldsymbol{\beta}}^* + \widetilde{\boldsymbol{Z}}'^\ell\tilde{\boldsymbol{\beta}}^* - \boldsymbol{X}'\tilde{\boldsymbol{\beta}}^*\|_2^2 \\
&\leq 2\|\widetilde{\boldsymbol{Z}}'^\ell\big(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\big)\|_2^2 + 2\|(\widetilde{\boldsymbol{Z}}'^\ell - \boldsymbol{X}')\tilde{\boldsymbol{\beta}}^*\|_2^2. \qquad \text{(S40)}
\end{aligned}$$

We shall bound the two terms on the right hand side of (S40) next.

Bounding $\|\widetilde{\boldsymbol{Z}}'^{\ell}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2$. Since $\widetilde{\boldsymbol{Z}}'^{\ell} = (1/\widehat{\rho}')\boldsymbol{Z}'^{\ell}$, we have

$$\begin{aligned}\|\widetilde{\boldsymbol{Z}}'^{\ell}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 &= \frac{1}{(\widehat{\rho}')^2}\|\boldsymbol{Z}'^{\ell}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\ &= \frac{1}{(\widehat{\rho}')^2}\|(\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}' + \rho\boldsymbol{X}')(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\ &\le \frac{2}{(\widehat{\rho}')^2}\|(\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}')(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 + 2\Big(\frac{\rho}{\widehat{\rho}'}\Big)^2\|\boldsymbol{X}'(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2. \quad \text{(S41)}\end{aligned}$$

Now, note that $\|\boldsymbol{Z}' - \boldsymbol{Z}'^{\ell}\|_2$ is the $(\ell+1)$-st largest singular value of $\boldsymbol{Z}'$. Therefore, by Weyl's inequality (Lemma 12), we have for any $\ell \ge r'$,

$$\|\boldsymbol{Z}' - \boldsymbol{Z}'^{\ell}\|_2 \le \|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2.$$

In turn, this gives

$$\|\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}'\|_2 \le \|\boldsymbol{Z}'^{\ell} - \boldsymbol{Z}'\|_2 + \|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2 \le 2\|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2.$$

Thus, we have

$$\|(\boldsymbol{Z}'^{\ell} - \rho\boldsymbol{X}')(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \le 4\|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2^2\ \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2. \quad \text{(S42)}$$

Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_\perp$ span the rowspace and nullspace of $\boldsymbol{X}$, respectively; similarly, recall that $\boldsymbol{H}'$ and $\boldsymbol{H}'_\perp$ are defined analogously with respect to $\boldsymbol{X}'$. As a result,

$$\begin{aligned}\|\boldsymbol{X}'(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 &= \|\boldsymbol{X}'(\boldsymbol{H} + \boldsymbol{H}_\perp)(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\ &\le 2\|\boldsymbol{X}'\boldsymbol{H}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 + 2\|\boldsymbol{X}'\boldsymbol{H}_\perp(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\ &\le 2\|\boldsymbol{X}'\|_2^2\ \|\boldsymbol{H}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 + 2\|\boldsymbol{X}'\|_2^2\ \|\boldsymbol{H}'\boldsymbol{H}_\perp(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2.\end{aligned}$$

Let $\widehat{\boldsymbol{H}}_r = \widehat{\boldsymbol{V}}_r\widehat{\boldsymbol{V}}_r^\top$ denote the projection matrix onto the rowspace of $\widetilde{\boldsymbol{Z}}^r$. Thus,

$$\begin{aligned}\|\boldsymbol{H}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 &= \|(\boldsymbol{H} - \widehat{\boldsymbol{H}}_r + \widehat{\boldsymbol{H}}_r)(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \\ &\le 2\|\boldsymbol{H} - \widehat{\boldsymbol{H}}_r\|_2^2\ \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 + 2\|\widehat{\boldsymbol{H}}_r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2.\end{aligned}$$

From (S9) and above, we obtain

$$\begin{aligned}\|\boldsymbol{H}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \le C\|\boldsymbol{H} - \widehat{\boldsymbol{H}}_r\|_2^2\ \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \\ + \frac{C}{\widehat{s}_r^2}\Big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\Big).\end{aligned}$$

Thus,

$$\begin{aligned}\|\boldsymbol{X}'(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \le C\|\boldsymbol{X}'\|_2^2\ \|\boldsymbol{H} - \widehat{\boldsymbol{H}}_r\|_2^2\ \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \\ + \frac{C\|\boldsymbol{X}'\|_2^2}{\widehat{s}_r^2}\Big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\Big)\end{aligned}$$

$$+ C\|\boldsymbol{X}'\|_2^2 \, \|\boldsymbol{H}'\boldsymbol{H}_\perp\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2. \tag{S43}$$

In summary, plugging (S42) and (S43) into (S41), we have

$$\begin{aligned}
\|\widetilde{\boldsymbol{Z}}'^{,\ell}(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*)\|_2^2 \le \frac{C}{(\widehat{\rho}')^2}\|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \\
+ C\Big(\frac{\rho}{\widehat{\rho}'}\Big)^2\|\boldsymbol{X}'\|_2^2 \, \|\boldsymbol{H} - \widehat{\boldsymbol{H}}_r\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \\
+ \frac{C\rho^2\|\boldsymbol{X}'\|_2^2}{(\widehat{\rho}')^2\widehat{s}_r^2}\Big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\Big) \\
+ C\Big(\frac{\rho}{\widehat{\rho}'}\Big)^2\|\boldsymbol{X}'\|_2^2 \, \|\boldsymbol{H}'\boldsymbol{H}_\perp\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2.
\end{aligned} \tag{S44}$$

Bounding $\|(\widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}')\tilde{\boldsymbol{\beta}}^*\|_2^2$. Using inequality (S1),

$$\|(\widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}')\tilde{\boldsymbol{\beta}}^*\|_2^2 \le \|\widetilde{\boldsymbol{Z}}'^{\ell} - \boldsymbol{X}'\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2. \tag{S45}$$

Combining. Incorporating (S44) and (S45) into (S40) with $\ell = r'$ yields

$$\|\widetilde{\boldsymbol{Z}}'^{r'}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}'\tilde{\boldsymbol{\beta}}^*\|_2^2 \le \Delta_1 + \Delta_2 + \Delta_3, \tag{S46}$$

where

$$\begin{aligned}
\Delta_1 &= \frac{C}{(\widehat{\rho}')^2}\|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 + C\left(\frac{\rho s_1'}{\widehat{\rho}'}\right)^2 \|\boldsymbol{H} - \widehat{\boldsymbol{H}}_r\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \\
&\quad + 2\|\boldsymbol{X}' - \widetilde{\boldsymbol{Z}}'^{r'}\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2, \\
\Delta_2 &= C\left(\frac{\rho s_1'}{\widehat{\rho}'\widehat{s}_r}\right)^2\Big(\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle\Big), \\
\Delta_3 &= C\left(\frac{\rho s_1'}{\widehat{\rho}'}\right)^2 \|\boldsymbol{H}'\boldsymbol{H}_\perp\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2.
\end{aligned}$$

Note that (S46) is a deterministic bound. We will now proceed to bound Δ_1 and Δ_2, first in high probability then in expectation.

Bound in high-probability. We first bound Δ_1. First we note that by adapting Lemma 10 with $\widehat{\rho}'$ in place of $\widehat{\rho}$, we obtain w.p. at least $1 - O(1/(mp)^{10})$,

$$\rho/2 \le \widehat{\rho}' \le \rho. \tag{S47}$$

By adapting Lemma 9 for $\boldsymbol{Z}', \boldsymbol{X}'$ in place of $\boldsymbol{Z}, \boldsymbol{X}$, we have w.p. at least $1 - O(1/(mp)^{10})$,

$$\|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2 \le C(K+1)(\gamma+1)(\sqrt{m} + \sqrt{p}).$$

Hence, using Theorem 4.1 and (S47), we have w.p. at least $1 - O(1/((n \wedge m)p)^{10})$

$$\frac{1}{(\widehat{\rho}')^2 m}\|\boldsymbol{Z}' - \rho\boldsymbol{X}'\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2$$

$$\leq C(K,\gamma,\sigma)\frac{\log(np)}{\rho^2}\left(1 \vee \frac{p}{m}\right)\cdot\left\{\frac{\|\tilde{\boldsymbol{\beta}}^*\|_2^2}{\mathsf{snr}^2} + \frac{\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1}{(n+p)\mathsf{snr}^2} + \frac{r\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{(n+p)\mathsf{snr}^2} + \frac{\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^4}\right\}, \tag{S48}$$

where $C(K,\gamma,\sigma) = C(K+1)^6(\gamma+1)^4(\sigma^2+1)$. Next, observe that $s_1' = O(\sqrt{mp})$, which follows from Assumption 3.3. Using this bound and recalling Lemma 2, (S18), and Theorem 4.1, it follows that w.p. at least $1 - O(1/(np)^{10})$

$$\begin{aligned}
&\left(\frac{\rho s_1'}{\widehat{\rho}'}\right)^2 \frac{1}{m}\|\boldsymbol{H} - \widehat{\boldsymbol{H}}_r\|_2^2 \, \|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 \\
&\quad\leq C(K,\gamma,\sigma)\log(np)\cdot\left\{\frac{p\|\tilde{\boldsymbol{\beta}}^*\|_2^2}{\mathsf{snr}^4} + \frac{\sqrt{np}\|\tilde{\boldsymbol{\beta}}^*\|_1}{(n+p)\mathsf{snr}^4} + \frac{rp\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{(n+p)\mathsf{snr}^4} + \frac{p\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^6}\right\} \\
&\quad\leq C(K,\gamma,\sigma)\log(np)\cdot\left\{\frac{p\|\tilde{\boldsymbol{\beta}}^*\|_2^2}{\mathsf{snr}^4} + \frac{\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1}{\mathsf{snr}^4} + \frac{r\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^4} + \frac{p\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^6}\right\}.
\end{aligned} \tag{S49}$$

Next, we adapt Lemma 3 for $\widetilde{\boldsymbol{Z}}', \boldsymbol{X}'$ in place of $\widetilde{\boldsymbol{Z}}, \boldsymbol{X}$ with $\ell = r'$. If $\rho \geq c(mp)^{-1}\log^2(mp)$, then w.p. at least $1 - O(1/(mp)^{10})$

$$\begin{aligned}
&\frac{1}{m}\|\boldsymbol{X}' - \widetilde{\boldsymbol{Z}}'^{r'}\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \\
&\quad\leq \frac{C(K,\gamma)}{m}\cdot\left\{\frac{(m+p)(m+\sqrt{m}\,\log(mp))}{\rho^4(s_r')^2} + \frac{r'+\sqrt{r'}\,\log(mp)}{\rho^2} + C\frac{\log(mp)}{\rho\, p}\right\}\|\tilde{\boldsymbol{\beta}}^*\|_1^2 \\
&\quad\leq \frac{C(K,\gamma)\log(mp)}{\rho^2}\cdot\left\{\frac{1}{\mathsf{snr}_{\text{test}}^2} + \frac{r'}{m}\right\}\|\tilde{\boldsymbol{\beta}}^*\|_1^2,
\end{aligned} \tag{S50}$$

where $C(K,\gamma) = C(K+1)^4(\gamma+1)^2$. Note that the above uses the inequality

$$\frac{m+p}{\rho^2(s_{r'}')^2} \leq \frac{1}{\mathsf{snr}_{\text{test}}^2},$$

which follows from the definition of $\mathsf{snr}_{\text{test}}^2$ in (10).

Next, we bound Δ_2. As per (S14), we have w.p. at least $1 - O(1/(np)^{10})$,

$$\begin{aligned}
&\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + \langle\widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle \\
&\leq C\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2\|\tilde{\boldsymbol{\beta}}^*\|_1^2 + C\sigma^2(r + \log(np)) + C\sigma\sqrt{n\log(np)}\|\tilde{\boldsymbol{\beta}}^*\|_1.
\end{aligned} \tag{S51}$$

Recalling Lemma 3 and the definition of snr, we have that w.p. at least $1 - O(1/(np)^{10})$,

$$\|\boldsymbol{X} - \widetilde{\boldsymbol{Z}}^r\|_{2,\infty}^2 \leq C(K,\gamma)\frac{\log(np)}{\rho^2}\cdot\left\{\frac{n}{\mathsf{snr}^2} + r\right\}. \tag{S52}$$

Using (S17), (S18), (S47), we have

$$\left(\frac{\rho s_1'}{\widehat{\rho}'\widehat{s}_r}\right)^2 \leq \frac{C(s_1')^2\rho^2}{\mathsf{snr}^2(n+p)}. \tag{S53}$$

Therefore, (S47), (S51), (S52), (S53), and the bound $s_1' = O(\sqrt{mp})$ altogether imply that w.p. at least $1 - O(1/((n \wedge m)p)^{10})$,

$$\begin{aligned}\frac{\Delta_2}{m} &\leq \frac{C(K,\gamma)(s_1')^2 \log(np)\rho^2}{m(n+p)\mathsf{snr}^2} \cdot \left\{ \frac{\sigma^2 r \|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\rho^2} + \sigma\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1 + \frac{n\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\rho^2\mathsf{snr}^2} \right\} \\ &\leq C(K,\gamma,\sigma)\log(np) \cdot \left\{ \frac{r\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^2} + \frac{\sqrt{n}\|\tilde{\boldsymbol{\beta}}^*\|_1}{\mathsf{snr}^2} + \frac{n\|\tilde{\boldsymbol{\beta}}^*\|_1^2}{\mathsf{snr}^4} \right\}, \end{aligned} \tag{S54}$$

where $C(K,\gamma,\sigma)$ is defined as in (S48). Moving on, we observe $\|\tilde{\boldsymbol{\beta}}^*\|_2 \leq \|\tilde{\boldsymbol{\beta}}^*\|_1$, and recall the assumption $\mathsf{snr} \geq C(K+1)(\gamma+1)$. With these in mind, we incorporate (S48), (S49), (S50), and (S54) into (S46) and simplify to establish

$$\begin{aligned}\frac{\Delta_1 + \Delta_2}{m} &\leq C(K,\gamma,\sigma)\log((n \vee m)p) \\ &\quad \cdot \left\{ \frac{\sqrt{n}}{\mathsf{snr}^2}\|\tilde{\boldsymbol{\beta}}^*\|_1 + \left(\frac{r\left(1 \vee \frac{p}{m}\right)}{\rho^2\mathsf{snr}^2} + \frac{r'}{\mathsf{snr}_{\text{test}}^2 \wedge m} + \frac{n \vee p}{\mathsf{snr}^4} \right) \|\tilde{\boldsymbol{\beta}}^*\|_1^2 \right\}. \end{aligned} \tag{S55}$$

Finally, we bound Δ_3. Following the arguments that led to (S49), we obtain w.p. at least $1 - O(1/(np)^{10})$

$$\frac{\Delta_3}{m} \leq C \cdot p \cdot \delta_\beta \cdot \|\boldsymbol{H}'\boldsymbol{H}_\perp\|_2^2. \tag{S56}$$

Combining (S55) and (S56) concludes the high-probability bound.

Bound in expectation. Here, we assume that $\{\langle \boldsymbol{x}_i, \boldsymbol{\beta}^* \rangle \in [-b, b] : i > n\}$. As such, we enforce $\{\widehat{y}_i \in [-b, b] : i > n\}$. With (S46), this yields

$$\text{MSE}_{\text{test}} \leq \frac{1}{m}\|\widetilde{\boldsymbol{Z}}'^{r'}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}'\tilde{\boldsymbol{\beta}}^*\|_2^2 \leq \frac{1}{m}\left(\Delta_1 + \Delta_2 + \Delta_3\right).$$

We define $\mathcal{E}$ as the event such that the bounds in (S47), (S48), (S49), (S50), (S52), and Lemma 4 hold. Thus, if $\mathcal{E}$ occurs, then combining (S48), (S49), (S50), and using the property $\|\tilde{\boldsymbol{\beta}}^*\|_2 \leq \|\tilde{\boldsymbol{\beta}}^*\|_1$ and assumption $\mathsf{snr} \geq C(K+1)(\gamma+1)$ gives

$$\begin{aligned}\frac{\mathbb{E}[\Delta_1|\mathcal{E}]}{m} \leq C(K,\gamma,\sigma)\log(n_{\max}p) \cdot \Bigg\{ &\frac{\sqrt{n}}{\mathsf{snr}^2}\left(\frac{1}{\mathsf{snr}^2} + \frac{1 \vee \frac{p}{m}}{\rho^2(n+p)} \right) \|\tilde{\boldsymbol{\beta}}^*\|_1 \\ &+ \left(\frac{r\left(1 \vee \frac{p}{m}\right)}{\rho^2\mathsf{snr}^2} + \frac{r'}{\mathsf{snr}_{\text{test}}^2 \wedge m} + \frac{n \vee p}{\mathsf{snr}^4} \right) \|\tilde{\boldsymbol{\beta}}^*\|_1^2 \Bigg\} \end{aligned} \tag{S57}$$

Next, we bound $\mathbb{E}[\Delta_2|\mathcal{E}]$. To do so, observe that $\boldsymbol{\varepsilon}$ is independent of the event $\mathcal{E}$. Thus, by (S35), we have

$$\begin{aligned}\mathbb{E}[\langle \widetilde{\boldsymbol{Z}}^r(\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*), \boldsymbol{\varepsilon}\rangle|\mathcal{E}] &= \mathbb{E}[\langle \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{U}}_r^\top \boldsymbol{X}\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle + \langle \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{U}}_r^\top\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle - \langle \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{S}}_r\widehat{\boldsymbol{V}}_r^\top\tilde{\boldsymbol{\beta}}^*, \boldsymbol{\varepsilon}\rangle|\mathcal{E}] \\ &= \mathbb{E}[\langle \widehat{\boldsymbol{U}}_r\widehat{\boldsymbol{U}}_r^\top\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\rangle|\mathcal{E}] \leq C\sigma^2 r. \end{aligned}$$

Combining the above inequality with (S52),

$$\frac{\mathbb{E}[\Delta_2|\mathcal{E}]}{m} \leq C(K,\gamma,\sigma)\log(np)\cdot\left\{\left(\frac{r}{\mathsf{snr}^2}+\frac{n}{\mathsf{snr}^4}\right)\|\tilde{\boldsymbol{\beta}}^*\|_1^2\right\}. \tag{S58}$$

Next, (S56) yields

$$\frac{\mathbb{E}[\Delta_3|\mathcal{E}]}{m} \leq C\cdot p\cdot\delta_\beta\cdot\|\boldsymbol{H}'\boldsymbol{H}_\perp\|_2^2. \tag{S59}$$

Due to truncation, observe that MSE_{test} is always bounded above by $4b^2$. Thus,

$$\begin{aligned}\mathbb{E}[\text{MSE}_{\text{test}}] &\leq \mathbb{E}[\text{MSE}_{\text{test}}|\mathcal{E}] + \mathbb{E}[\text{MSE}_{\text{test}}|\mathcal{E}^c]\ \mathbb{P}(\mathcal{E}^c) \\ &\leq \frac{1}{m}\mathbb{E}[\Delta_1+\Delta_2+\Delta_3|\mathcal{E}] + Cb^2\left(1/(np)^{10}+1/(mp)^{10}\right).\end{aligned} \tag{S60}$$

Plugging (S57), (S58), (S59) into (S60) and simplifying completes the proof.

Appendix E. Helpful Concentration Inequalities

In this section, we state and prove a number of helpful concentration inequalities used to establish our primary results.

Lemma 14 *Let X be a mean zero, sub-gaussian random variable. Then for any $\lambda\in\mathbb{R}$,*

$$\mathbb{E}\exp(\lambda X) \leq \exp\left(C\lambda^2\|X\|_{\psi_2}^2\right).$$

Lemma 15 *Let $X_1,\ldots,X_n$ be independent, mean zero, sub-gaussian random variables. Then,*

$$\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \leq C\sum_{i=1}^n\|X_i\|_{\psi_2}^2.$$

Theorem E.1 (Bernstein's inequality) *Let $X_1,\ldots,X_n$ be independent, mean zero, sub-exponential random variables. Then, for every $t\geq 0$, we have*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i\Big|\geq t\Big) \leq 2\exp\Big(-c\min\Big(\frac{t^2}{\sum_{i=1}^n\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i\|X_i\|_{\psi_1}}\Big)\Big),$$

where $c>0$ is an absolute constant.

Lemma 16 (Modified Hoeffding Inequality) *Let $\boldsymbol{X}\in\mathbb{R}^n$ be random vector with independent mean-zero sub-Gaussian random coordinates with $\|X_i\|_{\psi_2}\leq K$. Let $\boldsymbol{a}\in\mathbb{R}^n$ be another random vector that satisfies $\|\boldsymbol{a}\|_2\leq b$ almost surely for some constant $b\geq 0$. Then for all $t\geq 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n a_iX_i\Big|\geq t\Big) \leq 2\exp\Big(-\frac{ct^2}{K^2b^2}\Big),$$

where $c>0$ is a universal constant.

Proof Let $S_n = \sum_{i=1}^n a_i X_i$. Then applying Markov's inequality for any $\lambda > 0$, we obtain

$$\begin{aligned}\mathbb{P}\left(S_n \geq t\right) &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda t)\right) \\ &\leq \mathbb{E}\left[\exp(\lambda S_n)\right] \cdot \exp(-\lambda t) \\ &= \mathbb{E}_{\boldsymbol{a}}\left[\mathbb{E}\left[\exp(\lambda S_n) \mid \boldsymbol{a}\right]\right] \cdot \exp(-\lambda t).\end{aligned}$$

Now, conditioned on the random vector $\boldsymbol{a}$, observe that

$$\mathbb{E}\left[\exp(\lambda S_n)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp(\lambda a_i X_i)\right] \leq \exp\big(CK^2\lambda^2\|\boldsymbol{a}\|_2^2\big) \leq \exp\big(CK^2\lambda^2 b^2\big),$$

where the equality follows from conditional independence, the first inequality by Lemma 14, and the final inequality by assumption. Therefore,

$$\mathbb{P}\left(S_n \geq t\right) \leq \exp\big(CK^2\lambda^2 b^2 - \lambda t\big).$$

Optimizing over λ yields the desired result:

$$\mathbb{P}\left(S_n \geq t\right) \leq \exp\Big(-\frac{ct^2}{K^2 b^2}\Big).$$

Applying the same arguments for $-\langle \boldsymbol{X}, \boldsymbol{a}\rangle$ gives a tail bound in the other direction. ■

Lemma 17 (Modified Hanson-Wright Inequality) *Let $\boldsymbol{X} \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates with $\|X_i\|_{\psi_2} \leq K$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a random matrix satisfying $\|\boldsymbol{A}\|_2 \leq a$ and $\|\boldsymbol{A}\|_F^2 \leq b$ almost surely for some $a, b \geq 0$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\left|\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}]\right| \geq t\right) \leq 2 \cdot \exp\Big(-c\min\Big(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\Big)\Big).$$

Proof The proof follows similarly to that of Theorem 6.2.1 of Vershynin (2018). Using the independence of the coordinates of X, we have the following useful diagonal and off-diagonal decomposition:

$$\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}] = \sum_{i=1}^n \left(A_{ii}X_i^2 - \mathbb{E}[A_{ii}X_i^2]\right) + \sum_{i\neq j} A_{ij}X_iX_j.$$

Therefore, letting

$$p = \mathbb{P}\left(\boldsymbol{X}^\top \boldsymbol{A}X - \mathbb{E}[\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}] \geq t\right),$$

we can express

$$p \leq \mathbb{P}\Big(\sum_{i=1}^n \left(A_{ii}X_i^2 - \mathbb{E}[A_{ii}X_i^2]\right) \geq t/2\Big) + \mathbb{P}\Big(\sum_{i\neq j} A_{ij}X_iX_j \geq t/2\Big) =: p_1 + p_2.$$

We will now proceed to bound each term independently.

Step 1: diagonal sum. Let $S_n = \sum_{i=1}^n (A_{ii}X_i^2 - \mathbb{E}[A_{ii}X_i^2])$. Applying Markov's inequality for any $\lambda > 0$, we have

$$\begin{aligned} p_1 &= \mathbb{P}\left(\exp(\lambda S_n) \geq \exp(\lambda t/2)\right) \\ &\leq \mathbb{E}_{\boldsymbol{A}}\mathbb{E}\left[[\exp(\lambda S_n) \mid \boldsymbol{A}]\right] \cdot \exp(-\lambda t/2). \end{aligned}$$

Since the X_i are independent, sub-Gaussian random variables, $X_i^2 - \mathbb{E}[X_i^2]$ are independent mean-zero sub-exponential random variables, satisfying

$$\left\|X_i^2 - \mathbb{E}[X_i^2]\right\|_{\psi_1} \leq C\left\|X_i^2\right\|_{\psi_1} \leq C\|X_i\|_{\psi_2}^2 \leq CK^2.$$

Conditioned on $\boldsymbol{A}$ and optimizing over λ using standard arguments, yields

$$p_1 \leq \exp\Big(-c\min\Big(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\Big)\Big).$$

Step 2: off-diagonals. Let $S = \sum_{i\neq j} A_{ij}X_iX_j$. Again, applying Markov's inequality for any $\lambda > 0$, we have

$$p_2 = \mathbb{P}\left(\exp(\lambda S) \geq \exp(\lambda t/2)\right) \leq \mathbb{E}_{\boldsymbol{A}}\left[\mathbb{E}\left[\exp(\lambda S) \mid \boldsymbol{A}\right]\right] \cdot \exp(-\lambda t/2).$$

Let $\boldsymbol{g}$ be a standard multivariate gaussian random vector. Further, let $\boldsymbol{X}'$ and $\boldsymbol{g}'$ be independent copies of $\boldsymbol{X}$ and $\boldsymbol{g}$, respectively. Conditioning on $\boldsymbol{A}$ yields

$$\begin{aligned} \mathbb{E}\left[\exp(\lambda S)\right] &\leq \mathbb{E}\left[\exp\Big(4\lambda \boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}'\Big)\right] \qquad \text{(by Decoupling Remark 6.1.3 of Vershynin (2018))} \\ &\leq \mathbb{E}\left[\exp\Big(C_1\lambda \boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{g}'\Big)\right] \qquad \text{(by Lemma 6.2.3 of Vershynin (2018))} \\ &\leq \exp\Big(C_2\lambda^2\|\boldsymbol{A}\|_F^2\Big) \qquad \text{(by Lemma 6.2.2 of Vershynin (2018))} \\ &\leq \exp\big(C_2\lambda^2 b\big), \end{aligned}$$

where $|\lambda| \leq c/a$. Optimizing over λ then gives

$$p_2 \leq \exp\Big(-c\min\Big(\frac{t^2}{K^4 b}, \frac{t}{K^2 a}\Big)\Big).$$

Step 3: combining. Putting everything together completes the proof. ■

Appendix F. Proof of Theorem 6.1

Type I error. We first bound the Type I error, which anchors on Lemma 18, stated below. The proof of Lemma 18 can be found in Appendix F.2.

Lemma 18 *Suppose H_0 is true. Then,*

$$\begin{aligned} \widehat{\tau} = \|(\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{H})(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}')\|_F^2 \\ + 2\langle(\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}, (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\rangle_F. \end{aligned} \tag{S61}$$

We proceed to bound each term on the right-hand side of (S61) independently.

Bounding $\|(\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2$. By Lemma 2, we have w.p. at least $1 - \alpha_1$,

$$\|(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \leq \|\widehat{\boldsymbol{H}}^k - \boldsymbol{H}\|_2^2 \, \|\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \leq \frac{C\varsigma^2 r' \phi^2(\alpha_1)}{s_r^2}. \tag{S62}$$

Note that we have used the fact that $\|\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 = r'$.

Bounding $\|(\boldsymbol{I} - \boldsymbol{H})(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}')\|_F^2$. Observe that $(\boldsymbol{I} - \boldsymbol{H})$ is a projection matrix, and hence $\|\boldsymbol{I} - \boldsymbol{H}\| \leq 1$. By adapting Lemma 2, we have w.p. at least $1 - \alpha_2$

$$\|\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}'\|_F^2 \leq r' \|\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}'\|_2^2 \leq \frac{C\varsigma^2 r' (\phi'(\alpha_2))^2}{(s'_{r'})^2}. \tag{S63}$$

Note that we have used the following: (i) $\|\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}'\|_F = \|\sin\Theta\|_F$, where $\sin\Theta \in \mathbb{R}^{r' \times r'}$ is a matrix of principal angles between the two projectors (Absil et al., 2006), which implies $\operatorname{rank}(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}') \leq r'$; (ii) the standard norm inequality $\|\boldsymbol{A}\|_F \leq \sqrt{\operatorname{rank}(\boldsymbol{A})}\|\boldsymbol{A}\|_2$ for any matrix $\boldsymbol{A}$. Using the result above, we have

$$\|(\boldsymbol{I} - \boldsymbol{H})(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}')\|_F^2 \leq \|\boldsymbol{I} - \boldsymbol{H}\|_2^2 \, \|\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}'\|_F^2 \leq \frac{C\varsigma^2 r' (\phi'(\alpha_2))^2}{(s'_{r'})^2}. \tag{S64}$$

Bounding $\langle (\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell},\ (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell} \rangle_F$. Using the cyclic property of the trace operator, we have that

$$\begin{aligned} &\langle (\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell},\ (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell} \rangle_F = \operatorname{tr}\left((\widehat{\boldsymbol{H}}'^{\ell})^\top (\widehat{\boldsymbol{H}}^k - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\right) \\ &= \operatorname{tr}\left((\widehat{\boldsymbol{H}}^k - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\right). \end{aligned} \tag{S65}$$

Note that $\widehat{\boldsymbol{H}}^k - \boldsymbol{H}$ is symmetric, and $\boldsymbol{I} - \boldsymbol{H}$ and $\widehat{\boldsymbol{H}}'^{\ell}$ are both symmetric positive semidefinite (PSD). As a result, Lemmas 2 and 23 yield w.p. at least $1 - \alpha_1$

$$\begin{aligned} &\operatorname{tr}\left((\widehat{\boldsymbol{H}}^k - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\right) \leq \|\widehat{\boldsymbol{H}}^k - \boldsymbol{H}\|_2 \operatorname{tr}\left((\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\right) \\ &\leq \|\widehat{\boldsymbol{H}}^k - \boldsymbol{H}\|_2 \|\boldsymbol{I} - \boldsymbol{H}\|_2 \operatorname{tr}\left(\widehat{\boldsymbol{H}}'^{\ell}\right) \leq \frac{C\varsigma r' \phi(\alpha_1)}{s_r}. \end{aligned} \tag{S66}$$

Again, to arrive at the above inequality, we use $\|\boldsymbol{I} - \boldsymbol{H}\|_2 \leq 1$ and $\operatorname{tr}(\widehat{\boldsymbol{H}}'^{\ell}) = r'$.

Collecting terms. Collecting (S62), (S64), and (S66) with $\alpha_1 = \alpha_2 = \alpha/2$, w.p. at least $1 - \alpha$,

$$\widehat{\tau} \leq \frac{C\varsigma^2 r' \phi^2(\alpha/2)}{s_r^2} + \frac{C\varsigma^2 r' (\phi'(\alpha/2))^2}{(s'_{r'})^2} + \frac{C\varsigma r' \phi(\alpha/2)}{s_r}.$$

Defining the upper bound as $\tau(\alpha)$ completes the bound on the Type I error.

Type II error. Next, we bound the Type II error. We will leverage Lemma 19, the proof of which can be found in Appendix F.3.

Lemma 19 *The following equality holds:* $\widehat{\tau} = r' - c_1 - c_2$, *where*

$$\begin{aligned} c_1 &= \|\boldsymbol{H}\boldsymbol{H}^\top \boldsymbol{H}'\|_F^2 \\ c_2 &= \|(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^\ell\|_F^2 + \|\boldsymbol{H}(\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}')\|_F^2 \\ &\quad + 2\langle(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^\ell, \boldsymbol{H}\widehat{\boldsymbol{H}}'^\ell\rangle_F + 2\langle\boldsymbol{H}(\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}'), \boldsymbol{H}\boldsymbol{H}'\rangle_F. \end{aligned} \tag{S67}$$

We proceed to bound each term on the right hand side of (S67) separately.

Bounding $\|(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^\ell\|_F^2$. From (S62), we have that w.p. at least $1 - \alpha_1$,

$$\|(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^\ell\|_F^2 \leq \frac{C\varsigma^2 r'\phi^2(\alpha_1)}{s_r^2}. \tag{S68}$$

Bounding $\|\boldsymbol{H}(\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}')\|_F^2$. Using the inequality $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|_F$ for any two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, as well as the bound in (S63), we have w.p. at least $1 - \alpha_2$,

$$\|\boldsymbol{H}(\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}')\|_F^2 \leq \frac{C\varsigma^2 r'(\phi'(\alpha_2))^2}{(s'_{r'})^2}. \tag{S69}$$

Bounding $\langle(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^\ell, \boldsymbol{H}\widehat{\boldsymbol{H}}'^\ell\rangle_F$. Using an identical argument used to create the bounds in (S65) and (S66), but replacing $\boldsymbol{I} - \boldsymbol{H}$ with $\boldsymbol{H}$, we obtain w.p. at least $1 - \alpha_1$

$$\langle(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^\ell, \boldsymbol{H}\widehat{\boldsymbol{H}}'^\ell\rangle_F \leq \frac{C\varsigma r'\phi(\alpha_1)}{s_r}. \tag{S70}$$

Bounding $\langle\boldsymbol{H}(\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}'), \boldsymbol{H}\boldsymbol{H}'\rangle_F$. Like in the argument to produce the bound in (S66), we use Lemmas 2 and 23 to get that w.p. at least $1 - \alpha_2$,

$$\begin{aligned} &\langle\boldsymbol{H}(\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}'), \boldsymbol{H}\boldsymbol{H}'\rangle_F = \operatorname{tr}\left((\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}')\boldsymbol{H}^2\boldsymbol{H}'\right) \\ &= \operatorname{tr}\left((\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}')\boldsymbol{H}\boldsymbol{H}'\right) \leq \|\widehat{\boldsymbol{H}}'^\ell - \boldsymbol{H}'\|_2 \, \|\boldsymbol{H}\|_2 \operatorname{tr}\left(\boldsymbol{H}'\right) \leq \frac{C\varsigma r'\phi'(\alpha_2)}{s'_{r'}}. \end{aligned} \tag{S71}$$

Collecting terms. Combining (S68), (S69), (S70), (S71) with $\alpha_1 = \alpha_2 = \alpha/2$, and using the definition of $\tau(\alpha)$, we have that w.p. at least $1 - \alpha$,

$$c_2 \leq \tau(\alpha) + \frac{C\varsigma r'\phi'(\alpha/2)}{s'_{r'}}.$$

Hence, along with using Lemma 19, it follows that w.p. at least $1 - \alpha$,

$$\widehat{\tau} \geq r' - c_1 - \tau(\alpha) - \frac{C\varsigma r'\phi'(\alpha/2)}{s'_{r'}}. \tag{S72}$$

Now, suppose r' satisfies (14), which implies that H_1 must hold. Then, (S72) and (14) together imply $\mathbb{P}(\widehat{\tau} > \tau(\alpha)|H_1) \geq 1 - \alpha$. This completes the proof.

F.1 Proof of Corollary 6.1

Lemma 20 (Gaussian Matrices: Theorem 7.3.1 of Vershynin (2018)) *Let $\boldsymbol{A}$ be a $m \times n$ random matrix where the entries A_{ij} are Gaussian r.v.s with variance ς^2. Then for any $t > 0$, $\|\boldsymbol{A}\|_2 \leq \varsigma(\sqrt{m} + \sqrt{n} + t)$ w.p. at least $1 - 2\exp\big(-t^2\big)$.*

Lemma 21 *Let the setup of Lemma 2 hold. Further, assume the entries of $\boldsymbol{W}$ and $\boldsymbol{W}'$ are independent Gaussian r.v.s with variance ς^2. Then for any $\alpha \in (0, 1)$, we have w.p. at least $1 - \alpha$,*

$$\|\widehat{\boldsymbol{H}}^k - \boldsymbol{H}\|_2 \leq \frac{2\varsigma\phi(\alpha)}{s_r}, \quad \|\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}'\|_2 \leq \frac{2\varsigma\phi_{\text{post}}(\alpha)}{s'_{r'}}.$$

Proof The proof is identical to that of Lemma 2 except $\|\boldsymbol{Z} - \boldsymbol{X}\|$ is now bounded above using Lemma 20. ■

The remainder of the proof of Corollary 6.1 is identical to that of Theorem 6.1.

F.2 Proof of Lemma 18

Observe that

$$\begin{aligned}
\widehat{\tau} &= \|(\boldsymbol{I} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \\
&= \|(\boldsymbol{I} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell} - (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell} + (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \\
&= \|(\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell} + (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \\
&= \|(\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 + \|(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 + 2\langle(\boldsymbol{H} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}, (\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\rangle_F.
\end{aligned}$$

Under H_0, it follows that $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}' = 0$. As a result,

$$\begin{aligned}
\|(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 &= \|(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \\
&= \|(\boldsymbol{I} - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell} - (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}'\|_F^2 \\
&= \|(\boldsymbol{I} - \boldsymbol{H})(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}')\|_F^2.
\end{aligned}$$

Applying these two sets of equalities above together completes the proof.

F.3 Proof of Lemma 19

Because the columns of $\widehat{\boldsymbol{H}}'^{\ell}$ are orthonormal, $r' = \|\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 = \|\widehat{\boldsymbol{H}}^k\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 + \|(\boldsymbol{I} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2$. Therefore, it follows that

$$\widehat{\tau} = \|(\boldsymbol{I} - \widehat{\boldsymbol{H}}^k)\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 = r' - \|\widehat{\boldsymbol{H}}^k\widehat{\boldsymbol{H}}'^{\ell}\|_F^2. \tag{S73}$$

Now, consider the second term of the equality above.

$$\begin{aligned}
\|\widehat{\boldsymbol{H}}^k\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 &= \|\widehat{\boldsymbol{H}}^k\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell} + \boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \\
&= \|(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 + \|\boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 + 2\langle(\widehat{\boldsymbol{H}}^k - \boldsymbol{H})\widehat{\boldsymbol{H}}'^{\ell}, \boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell}\rangle_F.
\end{aligned} \tag{S74}$$

Further, analyzing the second term of (S74), we note that

$$\begin{aligned}\|\boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 &= \|\boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell}\|_F^2 \\ &= \|\boldsymbol{H}\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}\boldsymbol{H}' + \boldsymbol{H}\boldsymbol{H}'\|_F^2 \\ &= \|\boldsymbol{H}(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}')\|_F^2 + \|\boldsymbol{H}\boldsymbol{H}'\|_F^2 + 2\langle \boldsymbol{H}(\widehat{\boldsymbol{H}}'^{\ell} - \boldsymbol{H}'), \boldsymbol{H}\boldsymbol{H}'\rangle_F. \qquad \text{(S75)}\end{aligned}$$

Incorporating (S74) and (S75) into (S73), and recalling $c_1 = \|\boldsymbol{H}\boldsymbol{H}'\|_F^2 = \|\boldsymbol{H}\boldsymbol{H}^\top\boldsymbol{H}'\|_F^2$ completes the proof.

F.4 Helper Lemmas

Lemma 22 *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ be symmetric PSD matrices. Then,* $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \geq 0$.

Proof Let $\boldsymbol{B}^{1/2}$ denote the square root of $\boldsymbol{B}$. Since $\boldsymbol{A} \succeq 0$, we have

$$\begin{aligned}\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) &= \operatorname{tr}\Big(\boldsymbol{A}\boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2}\Big) \\ &= \operatorname{tr}\Big(\boldsymbol{B}^{1/2}\boldsymbol{A}\boldsymbol{B}^{1/2}\Big) \\ &= \sum_{i=1}^{n} (\boldsymbol{B}^{1/2}\boldsymbol{e}_i)'\boldsymbol{A}(\boldsymbol{B}^{1/2}\boldsymbol{e}_i) \geq 0.\end{aligned}$$

■

Lemma 23 *If $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is a symmetric PSD matrix, then* $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{A}) \cdot \operatorname{tr}(\boldsymbol{B})$, *where $\lambda_{\max}(\boldsymbol{A})$ is the top eigenvalue of $\boldsymbol{A}$.*

Proof Since $\boldsymbol{A}$ is symmetric, it follows that $\lambda_{\max}(\boldsymbol{A})\boldsymbol{I} - \boldsymbol{A} \succeq 0$. As a result, applying Lemma 22 yields $\operatorname{tr}((\lambda_{\max}(\boldsymbol{A})\boldsymbol{I} - \boldsymbol{A})\boldsymbol{B}) = \lambda_{\max}(\boldsymbol{A}) \cdot \operatorname{tr}(\boldsymbol{B}) - \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \geq 0$. ■

Appendix G. Towards a Lower Bound on Model Identification

We now take a first step towards establishing a lower bound on PCR's parameter estimation error in Lemma 24 below. Recall that Theorem 4.1 implies that PCR faithfully recovers the model parameter $\tilde{\boldsymbol{\beta}}^*$ provided snr grows sufficiently fast. Conversely, if $\mathsf{snr} = O(1)$, then Lemma 24 suggests the parameter estimation error is lower bounded by an absolute constant. To establish our result, we show that the Gaussian location model problem (Wu, 2020) is an instance of error-in-variables regression.

Lemma 24 *Let $n = O(p)$ and $\mathsf{snr} = O(1)$. Then,*

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\tilde{\boldsymbol{\beta}}^* \in \mathbb{B}_2} \mathbb{E}\|\widehat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}^*\|_2^2 = \Omega(1),$$

where $\mathbb{B}_2 = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\|_2 \leq 1\}$.

We make several important remarks. First and foremost, our result stated in Lemma 24 is only a partial correspondence with that stated in Theorem 4.1. The minimax bound in Lemma 24 is stated with $\rho = 1$, i.e., it does not capture the refined dependence on ρ. Meanwhile, (5) and (6) suggest that the error decays as ρ^{-4}. While this dependency on ρ may not be optimal, similar dependencies have appeared in error bounds within the error-in-variables literature, e.g., Loh and Wainwright (2012) and references therein. Establishing the optimal dependence with respect to ρ is interesting future work.

Moreover, Lemma 24 considers the constraint set $\mathbb{B}_2$, which contrasts with that considered in the main body of this work. Finally, as seen in the proof below, our reduction argument utilizes a specific choice of $\boldsymbol{X}$ while the main body of this work considers a fixed design matrix that the practitioner is unable to choose. Closing the gap on these limitations would significantly enhance the current lower bound, and we leave a formal treatment of this problem as important future work.

G.1 Proof of Lemma 24

Broadly, we proceed in three steps: (i) stating the Gaussian location model (GLM) and an associated minimax result; (ii) reducing GLM to an instance of error-in-variables regression; (iii) establishing a minimax result on the parameter estimation error of error-in-variables using the GLM minimax result.

Gaussian location model. Below, we introduce the GLM setting through a well-known minimax result.

Lemma 25 (Theorem 12.4 of Wu (2020)) *Let $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}^*, \sigma^2 \boldsymbol{I}_p)$, where $\boldsymbol{I}_p \in \mathbb{R}^{p \times p}$ is the identity matrix and $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \mathbb{R}^p$. Given $\boldsymbol{\theta}$, let $\widehat{\boldsymbol{\theta}}$ be any estimator of $\boldsymbol{\theta}^*$. Then,*

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \mathbb{B}_2} \mathbb{E} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 = \Theta(\sigma^2 p \wedge 1).$$

Reducing GLM to error-in-variables. We will now show how an instance of GLM can be reduced to an instance of error-in-variables. Towards this, we follow the setup of Lemma 25 and define $\boldsymbol{\beta}^* = \boldsymbol{\theta}^*$, $\boldsymbol{\beta} = \boldsymbol{\theta}$, and $s = 1/\sigma$. For convenience, we write $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{\eta}$, where the entries of $\boldsymbol{\eta}$ are independent Gaussian r.v.s with mean zero and variance $1/s^2$; hence $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}^*, (1/s^2)\boldsymbol{I}_p)$. Now, recall that the error-in-variables setting reveals a response vector $y = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and covariate $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{W}$, where the parameter estimation objective is to recover $\boldsymbol{\beta}^*$ from $(\boldsymbol{y}, \boldsymbol{Z})$. Below, we construct instances of these quantities using $\boldsymbol{\beta}, \boldsymbol{\beta}^*$ as follows:

(i) Let the SVD of $\boldsymbol{X}$ be defined as $\boldsymbol{X} = s\boldsymbol{u} \otimes \boldsymbol{v}$, where $\boldsymbol{u} = (1, 0, \ldots, 0)^T \in \mathbb{R}^n$ and $\boldsymbol{v} = \boldsymbol{\beta}^*$. Note by construction, $\text{rank}(\boldsymbol{X}) = 1$ and $\boldsymbol{\beta}^* \in \text{rowspan}(\boldsymbol{X})$.

(ii) To construct $\boldsymbol{y}$, we first sample $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ whose entries are independent standard normal r.v.s. Next, we define $\boldsymbol{y} = s\boldsymbol{u} + \boldsymbol{\varepsilon}$. From (i), we note that $\boldsymbol{X}\boldsymbol{\beta}^* = s\boldsymbol{u}$ such that $\boldsymbol{y}$ can be equivalently expressed as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$.

(iii) Let $\boldsymbol{Z} = s\boldsymbol{u} \otimes \boldsymbol{\beta}$. By construction, it follows that $\boldsymbol{Z} = \boldsymbol{X} + s\boldsymbol{u} \otimes \boldsymbol{\eta}$. Note that $\boldsymbol{W} = s\boldsymbol{u} \otimes \boldsymbol{\eta}$ is an $n \times p$ matrix whose entries in the first row are independent standard normal r.v.s and the remaining entries are zero.

Establishing minimax result. As stated above, the error-in-variables parameter estimation task is to construct $\widehat{\boldsymbol{\beta}}$ from $(\boldsymbol{y}, \boldsymbol{Z})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ vanishes as n, p grow. Using the above reduction combined with Lemma 25, it follows that

$$\inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^* \in \mathbb{B}_2} \mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 = \Theta(p/s^2 \wedge 1).$$

To attain our desired result, it suffices to establish that $p/s^2 = \Omega(1)$. By (5) and under the assumption $n = O(p)$, we have that $s^2 \leq 2\mathsf{snr}^2(n+p) \leq c\mathsf{snr}^2 p$ for some $c > 0$. As such, if $\mathsf{snr} = O(1)$, then the minimax error is bounded below by a constant.

References

A. Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021. doi: 10.1257/jel.20191450. URL https://www.aeaweb.org/articles?id=10.1257/jel.20191450.

A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 2003.

A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American Statistical Association*, 2010.

P.-A. Absil, A. Edelman, and P. Koev. On the largest principal angle between random subspaces. *Linear Algebra and its Applications*, 414(1):288 – 294, 2006. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2005.10.004. URL http://www.sciencedirect.com/science/article/pii/S0024379505004878.

A. Agarwal, D. Shah, D. Shen, and D. Song. On robustness of principal component regression. *Journal of the American Statistical Association*, 2021.

A. Agarwal, A. Agarwal, and S. Vijaykumar. Synthetic combinations: A causal inference framework for combinatorial interventions. In *Advances in Neural Information Processing Systems*, volume 36, pages 19195–19216, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3d17b7f7d52c83ab6e97e2dc0bda2e71-Paper-Conference.pdf.

M. Amjad, V. Mishra, D. Shah, and D. Shen. mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2), 2019.

M. J. Amjad, D. Shah, and D. Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19:1–51, 2018.

S. Athey and G. W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, May 2017. doi: 10.1257/jep.31.2.3. URL https://www.aeaweb.org/articles?id=10.1257/jep.31.2.3.

J. Bai. Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279, 2009. ISSN 00129682, 14680262.

J. Bai and S. Ng. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021. doi: 10.1080/01621459.2021.1967163. URL https://doi.org/10.1080/01621459.2021.1967163.

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1907378117.

A. Belloni, V. Chernozhukov, A. Kaul, M. Rosenbaum, and A. B. Tsybakov. Pivotal estimation via self-normalization for high-dimensional linear models with errors in variables. *arXiv:1708.08353*, 2017a.

A. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming approaches to high-dimensional errors-in-variables models. *Journal of the Royal Statistical Society*, 79:939–956, 2017b.

E. Ben-Michael, A. Feller, and J. Rothstein. Synthetic controls with staggered adoption. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):351–381, 12 2021. ISSN 1369-7412. doi: 10.1111/rssb.12448. URL https://doi.org/10.1111/rssb.12448.

S. Bhattacharya and S. Chatterjee. Matrix completion with data-dependent missingness probabilities. *IEEE Transactions on Information Theory*, 68(10):6762–6773, 2022. doi: 10.1109/TIT.2022.3170244.

C. M. Bishop. Bayesian pca. In *Advances in neural information processing systems*, pages 382–388, 1999.

C. Cai, G. Li, H. V. Poor, and Y. Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/a1519de5b5d44b31a01de013b9b51a80-Paper.pdf.

T. T. Cai and P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159 – 2179, 2006. doi: 10.1214/009053606000000830. URL https://doi.org/10.1214/009053606000000830.

M. D. Cattaneo, Y. Feng, F. Palomba, and R. Titiunik. Uncertainty quantification in synthetic controls with staggered treatment adoption, 2024. URL https://arxiv.org/abs/2210.05026.

R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, pages 245–276, 1966.

G. Chao, Y. Luo, and W. Ding. Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction*, 1(1):341–358, 2019. ISSN 2504-4990. doi: 10.3390/make1010020. URL http://www.mdpi.com/2504-4990/1/1/20.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43:177–214, 2015.

Y. Chen and C. Caramanis. Orthogonal matching pursuit with noisy and missing data: Low and high dimensional results. *arXiv preprint arXiv:1206.0823*, 2012.

Y. Chen and C. Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *International Conference on Machine Learning*, pages 383–391, 2013.

A. Datta and H. Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426, 2017.

M.-H. Descary, V. M. Panaretos, et al. Functional data analysis by matrix completion. *Annals of Statistics*, 47(1):1–38, 2019.

J. Fan, W. Wang, and Y. Zhong. An ℓ_∞ eigenvector perturbation bound and its application. *Journal of Machine Learning Research*, 18(207):1–42, 2018. URL http://jmlr.org/papers/v18/16-140.html.

M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is. *IEEE Transactions on Information Theory*, 60(8):5040–5053, Aug 2014. ISSN 1557-9654. doi: 10.1109/tit.2014.2323359. URL http://dx.doi.org/10.1109/TIT.2014.2323359.

Z. Guo, D. Ćevid, and P. Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *The Annals of Statistics*, 50, 06 2022. doi: 10.1214/21-AOS2152.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70 – 91, 2007. doi: 10.1214/009053606000000957. URL https://doi.org/10.1214/009053606000000957.

P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493 – 1517, 2006. doi: 10.1214/009053606000000272. URL https://doi.org/10.1214/009053606000000272.

P. Hoff. Model averaging and dimension selection for the singular value decomposition. *Journal of the American Statistical Association*, 102:674–685, 02 2007. doi: 10.2307/27639896.

I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society*, 31(3):300–303, 1982.

A. Kaul and H. L. Koul. Weighted ℓ_1-penalized corrected quantile regression for high dimensional measurement error models. *Journal of Multivariate Analysis*, 140:72–91, 2015.

Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321 – 3351, 2010. doi: 10.1214/10-AOS813. URL https://doi.org/10.1214/10-AOS813.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

P.-l. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

W. Ma and G. H. Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In *Advances in Neural Information Processing Systems*, 2019.

Y. F. Matias D. Cattaneo and R. Titiunik. Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021. doi: 10.1080/01621459.2021.1979561. URL https://doi.org/10.1080/01621459.2021.1979561. PMID: 35756161.

J. Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Master's Thesis*, 1923.

A. B. Owen and P. O. Perry. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *The Annals of Applied Statistics*, 3(2):564 – 594, 2009. doi: 10.1214/08-AOAS227. URL https://doi.org/10.1214/08-AOAS227.

S. Roman. Graduate texts in mathematics: Advanced linear algebra. *Springer*, 2008.

M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix estimation. *The Annals of Statistics*, 38(5):2620–2651, 2010.

M. Rosenbaum and A. B. Tsybakov. Improved matrix uncertainty selector. *From Probability to Statistics and Back: High-Dimensional Models and Processes*, 9:276–290, 2013.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444.

T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, volume 48, pages 1670–1679, 20–22 Jun 2016. URL http://proceedings.mlr.press/v48/schnabel16.html.

A. M. Shaikh and P. Toulis. Randomization tests in observational studies with staggered adoption of treatment. *Journal of the American Statistical Association*, 116(536):1835–1848, 2021. doi: 10.1080/01621459.2021.1974458. URL https://doi.org/10.1080/01621459.2021.1974458.

J. Shao and X. Deng. Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40(2):812 – 831, 2012. doi: 10.1214/12-AOS982. URL https://doi.org/10.1214/12-AOS982.

A. Sportisse, C. Boyer, and J. Josses. Estimation and imputation in probabilistic principal component analysis with missing not at random data. *Advances in Neural Information Processing Systems*, 33, 2020.

C. Squires, D. Shen, A. Agarwal, D. Shah, and C. Uhler. Causal imputation via synthetic interventions. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 688–711. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/squires22b.html.

G. Strang. Linear algebra and its applications. *Brooks/Cole Cengage Learning*, 2006.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

M. Udell and A. Townsend. Nice latent variable models have log-rank. *ArXiv*, abs/1705.07474, 2017.

M. Udell and A. Townsend. Why are big data matrices approximately low rank?, 2018.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978. ISSN 00401706.

Y. Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics, January 2020.

J. Xu. Rates of convergence of spectral methods for graphon estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5433–5442, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/xu18a.html.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873 – 2903, 2005. doi: 10.1214/009053605000000660. URL https://doi.org/10.1214/009053605000000660.

D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21(232):1–41, 2020. URL http://jmlr.org/papers/v21/19-545.html.