On Robustness of Principal Component Regression

Citation: Agarwal, Anish, Shah, Devavrat, Shen, Dennis and Song, Dogyoon. 2021. "On Robustness of Principal Component Regression." Journal of the American Statistical Association, 116 (536).

As Published: 10.1080/01621459.2021.1928513

Publisher: Informa UK Limited

Persistent URL: https://hdl.handle.net/1721.1/143874

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

On Robustness of Principal Component Regression

Anish Agarwal, Devavrat Shah, Dennis Shen, Dogyoon Song
MIT

Abstract

Principal component regression (PCR) is a simple, but powerful and ubiquitously utilized method. Its effectiveness is well established when the covariates exhibit low-rank structure. However, its ability to handle settings with noisy, missing, and mixed-valued, i.e., discrete and continuous, covariates is not understood and remains an important open challenge. As the main contribution of this work we establish the robustness of PCR, without any change, in this respect and provide meaningful finite-sample analysis.

To do so, we establish that PCR is equivalent to performing linear regression after pre-processing the covariate matrix via hard singular value thresholding (HSVT). As a result, in the context of counterfactual analysis using observational data, we show PCR is equivalent to the recently proposed robust variant of the synthetic control method, known as robust synthetic control (RSC). As an immediate consequence, we obtain finite-sample analysis of the RSC estimator that was previously absent. As an important contribution to the synthetic controls literature, we establish that an (approximate) linear synthetic control exists in the setting of a generalized factor model, or latent variable model; traditionally in the literature, the existence of a synthetic control needs to be assumed to exist as an axiom. We further discuss a surprising implication of the robustness property of PCR with respect to noise, i.e., PCR can learn a good predictive model even if the covariates are tactfully transformed to preserve differential privacy.

Finally, this work advances the state-of-the-art analysis for HSVT by establishing stronger guarantees with respect to the $\ell_{2,\infty}$-norm rather than the frobenius norm as is commonly done in the matrix estimation literature, which may be of interest in its own right.

Keywords: principal component regression, synthetic controls, error-in-variables regression, hard singular value thresholding, matrix estimation

1 Introduction

A common thread of many modern datasets is that they are high-dimensional, and often noisy and partially observed. When such datasets are used for regression, this means that *both* the response variables (also known as the label of target) and the covariates (also known as features) are corrupted. This setting is known in the statistics literature as error-in-variables regression. Another common feature of most real-world datasets are that they are mixed valued, i.e., contain both discrete and continuous data, which further complicates the regression procedure. Within this context, we are interested in developing a better understanding of a popular prediction method known as principal component regression (PCR). Indeed, PCR's ability to handle settings with noisy, missing, and mixed-valued covariates is not understood and remains an important open challenge [17].

A further motivation of this work is to connect the error-in-variables setting to the exciting and growing literature on synthetic controls (SC), a standard framework in econometrics (and beyond) to make counterfactual predictions utilizing only observational data ([2, 1, 24, 22, 49, 8, 7, 5, 30, 29, 25, 13, 6]). Broadly speaking, there is a notion of a "target" and "donor" units, for which we collect observations over time. While the donors units remain under control, the target undergoes an intervention at some time period. Here, the goal is to estimate what would have happened to the target unit had it also remained under control. Towards answering this question, standard SC methods build a synthetic model of the target unit using observations associated with the donor units. In the language of regression, the target unit observations represent the response variables and the donor unit observations represent the covariates. In the SC literature, the observations associated with both the target and donor units are assumed to be *noisily* observed due to the presence of idiosyncratic shocks at each time step. As a result, SC can be seen as an instance of error-in-variables regression; more generally, panel data settings, where one collects measurements over time, can also be viewed through this error-in-variables lens.

As the main contribution of this work, we establish the effectiveness of PCR, without any change, for error-in-variables regression and provide meaningful finite-sample analysis for both in- and out-of-sample prediction error. Given the connection between error-in-variables regression and SC, our analysis also implies that using PCR in this context leads to it implicitly de-noising the observations we have of the donor units, which are corrupted by idiosyncratic shocks. Thus, we advocate for PCR's usage in panel data settings.

1.1 Problem Statement

In a typical prediction problem setup, we are given access to a labeled dataset $\{(Y_i, \boldsymbol{A}_{i,\cdot})\}$ over $i \geq 1$; here, $Y_i \in \mathbb{R}$ represents the response variable we wish to predict, and $\boldsymbol{A}_{i,\cdot} \in \mathbb{R}^{1\times p}$ represents the associated covariate to be utilized in the prediction process. Let $N \geq 1$ denote the total number of observations, where the number of predictors p can possibly exceed N. Let $\boldsymbol{A} \in \mathbb{R}^{N\times p}$ denote the matrix of true covariates.

Error-in-variables. Rather than perfectly observing the covariates $\boldsymbol{A}$, the error-in-variables setting only reveals a corrupted version denoted as $\boldsymbol{Z} \in \mathbb{R}^{N\times p}$. That is, the (i,j)-th entry of $\boldsymbol{Z}$, denoted as Z_{ij}, is defined as $A_{ij} + \eta_{ij}$ with probability ρ and $\star$ with probability $1-\rho$, for some $\rho \in (0,1]$; here, $\star$ denotes a missing value and η_{ij} denotes the noise in the (i,j)-th entry. In other words, each entry Z_{ij} is observed with probability ρ, independently of other entries; however, even when observed, Z_{ij} is still only a noisy instance of the true A_{ij}.

Approximate linear model. We assume the response variables are generated as follows: for $i \in [N]$, the random response Y_i is associated with the covariate $\boldsymbol{A}_{i,\cdot}$ via

$$Y_i = \boldsymbol{A}_{i,\cdot}\beta^* + \epsilon_i + \phi_i, \tag{1}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown latent model parameter, $\epsilon_i \in \mathbb{R}$ denotes zero mean response noise with variance bounded by σ^2, and $\phi_i \in \mathbb{R}$ is the linear model misspecification, or mismatch, error; for simplicity, we assume the mismatch error is deterministic. Additionally, the observed response variables Y_i are restricted to a subset of the N observations. More formally, we denote $\Omega \subset [N]$, with $|\Omega| = n < N$, as the index set of observed responses, i.e., we observe Y_i for $i \in \Omega$.

Goal. Given noisy observations of all N covariates $\{\boldsymbol{Z}_{1,\cdot},\ldots,\boldsymbol{Z}_{N,\cdot}\}$ and a subset of response variables $\{Y_i : i \in \Omega\}$, our aim is to produce an estimate $\widehat{Y} \in \mathbb{R}^N$ so that the prediction error is minimized. Specifically, we measure performance in terms of the *training error*

$$\text{MSE}_{\Omega}(\widehat{Y}) = \frac{1}{n}\mathbb{E}\left[\sum_{i\in\Omega}(\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*)^2\right] \tag{2}$$

and *testing error*

$$\text{MSE}(\widehat{Y}) = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}(\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*)^2\right]. \tag{3}$$

We note that for the bound $\text{MSE}(\widehat{Y})$ to be meaningful, $|\Omega^c| = N - n$ (the size of the test set) should be of the same order as that of the the training set $|\Omega| = n$.

Transductive semi-supervised learning. It is worth remarking that in (3), the algorithm is given access to the observations associated with the covariates for *both* training and testing data during the training procedure. Of course, however, the algorithm does *not* access the test response variables. This is commonly referred to in the literature as transductive semi-supervised learning; here, we want to infer the response variables for the specific unlabeled data. Traditionally, it is assumed that a statistical estimator only has access to the training covariates and response variables during the model learning process. The reason we consider a transductive learning setting is a consequence of the nature of the algorithm of interest, PCR. Specifically, PCR pre-processes the covariates using PCA, which changes the training procedure if only a subset of the covariates are utilized. Therefore, to allow for a meaningful evaluation, it is natural to allow the algorithm to have access to *all* available covariate information. Indeed, as we will discuss in Section 4, as well as Appendices B and C, it is natural to have access to all covariates in many important real-world applications.

1.2 Contributions

PCR implicitly de-noises. As the main contribution of this work, we argue that PCR, without any change, is robust to noise and missing values in the observed covariates. In particular, despite only having access to $\boldsymbol{Z}$, we show the training error of PCR scales (up to logarithmic factors) as $\rho^{-4}r/\min(n,p)+\|\phi\|_2^2/n$, where ρ denotes the fraction of observed (noisy) covariates and r is the rank of $\boldsymbol{A}$ (Corollary 3.1). That is, PCR *implicitly de-noises* $\boldsymbol{Z}$ by projecting it onto the subspace spanned by the top r right singular vectors. We note that the prediction error rate of r/n for the training data matches (up to log factors) the minimax rate achievable by ordinary least squares (OLS) if one had perfectly observed the true underlying covariate matrix $\boldsymbol{A}$ (see [46] and references therein).

We extend our results to the case where $\boldsymbol{A}$ is only approximately low-rank (Theorem 3.1 and Corollaries 3.2 and 3.3). To the best of our knowledge, under this setting, there do not exist prediction consistency results for OLS or regularized variants thereof such as Lasso and Ridge, without making additional assumptions on the sparsity of β^*. This remains true even if $\boldsymbol{A}$ is perfectly observed. Thus, the first step in PCR of finding a low-dimensional representation is likely crucial for this setting, and further motivated if the covariates are noisily observed. Given the ubiquity of approximately low-rank matrices in real-world datasets, it reinforces the utility and robustness of applying PCR in practice.

Moreover, we note that PCR does *not* require any knowledge about the underlying noise model that corrupts the covariates in order to to have vanishing train and test errors. Despite the exciting recent advancement in the high-dimensional error-in-variables literature, such as in [31, 20, 38], the current inventory of methods require knowledge of the underlying covariate noise model (in particular, exact knowledge of its second moment of matrix) and do not provide finite sample guarantees for train or test error. We do note, however, that the aim of these previous papers is to estimate the latent linear model parameter β^* (assuming it is sparse), rather than to analyze prediction errors. For a detailed comparison, see Appendix A.

PCR implicitly regularizes. We define an appropriate notion of generalization error for the transductive learning setting we consider. We establish that the testing prediction error of PCR is bounded above by the training error plus a term that scales as $k^{5/2}/\sqrt{n}$, where k is the number of retained principal components (Theorem 3.2). Our testing error result provides a systematic way to select the correct number of principal components in a data-driven manner, i.e., to choose the value of k that minimizes the training error plus the generalization penalty term $k^{5/2}/\sqrt{n}$.

Our test error analysis utilizes the standard framework of Rademacher complexity (see [10] and references therein). However, there are two crucial differences that we need to overcome in order to obtain sharp, meaningful bounds. First, our notion of generalization is different from that of the traditional setup since the noisy test covariates (but not responses) are included in the training process, which requires careful analysis. Second, we argue that the Rademacher complexity under PCR scales with the dimensionality of the number of principle components utilized, denoted as k, rather than the ambient covariate dimension p. To do so, we identify the Rademacher complexity class of PCR with k-sparse β's.

PCR applications. We discuss the robustness of PCR to contaminated covariates by analyzing its ability to learn a predictive model when only differentially private covariates are available. In particular, we find that it is feasible for PCR to achieve good prediction accuracy and simultaneously maintain differential privacy of the covariates (Appendix B). We also describe how the robustness of PCR allows it to seamlessly utilize mixed valued covariates under a general probabilistic model (Appendix C).

SC literature. First, we note that regardless of method used to construct synthetic controls, the fundamental hypothesis that drives these prior works is the existence of a linear relationship between the target and donors; in fact, the original proposal of [2, 1] suggests restricting the linear model coefficients to be non-negative and sum to one, i.e., a convex combination. However, it is not clear when such a hypothesis holds. Second, meaningful finite-sample analysis of the mean-squared post-intervention error of SC has remained elusive. We tackle these two questions via our results on PCR. Towards the first question, we establish that (approximate) synthetic controls exist under a generalized factor model (also known as a latent variable model). Here, the measurement associated with a given unit and time is a sufficiently smooth function of the latent unit and time factors. Therefore in a general sense, a synthetic control almost always exists and need not be assumed as a hypothesis or axiom (see Proposition 4.1). Towards the second question, we show that PCR is identical to a recently proposed SC estimator known as robust synthetic control (RSC) [5]. Hence, we immediately establish meaningful training (pre-intervention) and testing (post-intervention) error guarantees for RSC (see Theorem 4.1).

1.3 Organization of Paper

In Section 2, we describe the PCR algorithm. Section 3 then details the various training and test prediction error bounds for PCR and the conditions under which they hold. In Section 4, we formally connect PCR to SC. In Appendix A, we do a detailed comparison with previous related works. In Appendices B and C, we discuss the application of PCR for differentially private regression and mixed valued covariates, respectively. The remaining appendices are to prove our theoretical results.

2 Principal Component Regression

We recall the description of PCR, as in [26]. We suggest a minor modification of PCR in the presence of missing data where we simply re-scale the observed covariates by the inverse of the fraction of observed data.

Algorithm. Let $\widehat{\rho}$ denote the fraction of observed entries in $\boldsymbol{Z}$, i.e., $\widehat{\rho} = 1/(Np)\sum_{i=1}^{N}\sum_{j=1}^{p}\mathbb{1}(Z_{ij} \neq \star) \vee 1/(Np)$. Let $\widetilde{\boldsymbol{Z}} \in \mathbb{R}^{N\times p}$ represent the rescaled version of $\boldsymbol{Z}$, where every unobserved value $\star$ is replaced by 0, i.e., $\widetilde{Z}_{ij} = Z_{ij}/\widehat{\rho}$ if $Z_{ij} \neq \star$ and 0 otherwise.

The singular value decomposition (SVD) of $\widetilde{\boldsymbol{Z}}$ is denoted as $\widetilde{\boldsymbol{Z}} = \boldsymbol{USV}^T = \sum_{i=1}^{N} s_i u_i v_i^T$, where

$\boldsymbol{U} \in \mathbb{R}^{N\times N}$, $\boldsymbol{S} \in \mathbb{R}^{N\times p}$, and $\boldsymbol{V} \in \mathbb{R}^{p\times p}$. Without loss of generality, assume that the singular values s_i's are arranged in decreasing order, i.e., $s_1 \geq ... \geq s_N \geq 0$. Note that $\boldsymbol{U} = [u_1,...,u_N]$ and $\boldsymbol{V} = [v_1,...,v_p]$ are orthogonal matrices, i.e., the u_i's and v_j's are orthonormal vectors.

For any $k \in [N]$, let $\boldsymbol{U}_k = [u_1,...,u_k]$, $\boldsymbol{V}_k = [v_1,...,v_k]$, and $\boldsymbol{S}_k = \text{diag}(s_1,...,s_k)$. Then, the k-dimensional representation of $\widetilde{\boldsymbol{Z}}$, as per PCA, is given by $\boldsymbol{Z}^{\text{PCR},k} = \widetilde{\boldsymbol{Z}}\boldsymbol{V}_k$. Let $\beta^{\text{PCR},k} \in \mathbb{R}^k$ be the solution to the linear regression problem under $\boldsymbol{Z}^{\text{PCR},k}$, i.e., $\beta^{\text{PCR},k}$ is the minimizer of

$$\mathsf{minimize} \sum_{i\in\Omega} \left(Y_i - \boldsymbol{Z}_{i\cdot}^{\text{PCR},k} w\right)^2 \text{ over } w \in \mathbb{R}^k.$$

Then, the estimated N-dimensional response vector $\widehat{Y}^{\text{PCR},k} = \boldsymbol{Z}^{\text{PCR},k}\beta^{\text{PCR},k}$.

Intuition. Using all the noisily observed observed covariates, PCR first finds a k dimensional representation of the covariate matrix using the method of principal component analysis (PCA), where k might be much smaller than p. Specifically, PCA projects every covariate $\boldsymbol{Z}_{i,\cdot}$ onto the subspace spanned by the top k right singular vectors of the observed covariate matrix, $\boldsymbol{Z}$. PCR then uses the k-dimensional features to perform linear regression.

2.1 Connecting PCR to the Matrix Estimation Literature

To establish our results, we study PCR via its equivalence with performing linear regression after pre-processing covariates via hard singular value thresholding (HSVT), as described below.

Linear regression with covariate pre-processing via HSVT. Given any $\lambda > 0$, we define the map $\text{HSVT}_\lambda : \mathbb{R}^{N\times p} \to \mathbb{R}^{N\times p}$, which simply shaves off the input matrix's singular values that are below the threshold λ. Precisely, given a matrix $\boldsymbol{B} \in \mathbb{R}^{N\times p}$, denote its SVD as $\boldsymbol{B} = \sum_{i=1}^N \sigma_i x_i y_i^T$, and let $\text{HSVT}_\lambda(\boldsymbol{B}) = \sum_{i=1}^N \sigma_i \mathbb{1}(\sigma_i \geq \lambda) x_i y_i^T$. For any $k \in [N]$, given $\widetilde{\boldsymbol{Z}}$ as before, define $\boldsymbol{Z}^{\text{HSVT},k} = \text{HSVT}_{s_k}(\widetilde{\boldsymbol{Z}})$.

Let $\beta^{\mathrm{HSVT},k} \in \mathbb{R}^p$ be a solution of linear regression under $\boldsymbol{Z}^{\mathrm{HSVT},k}$, i.e., $\beta^{\mathrm{HSVT},k}$ is the minimizer of

$$\textsf{minimize} \sum_{i\in\Omega}\Big(Y_i - \boldsymbol{Z}_{i\cdot}^{\mathrm{HSVT},k} w\Big)^2 \text{ over } w \in \mathbb{R}^p.$$

Then, the estimated N-dimensional response vector $\widehat{Y}^{\mathrm{HSVT},k} = \boldsymbol{Z}^{\mathrm{HSVT},k}\beta^{\mathrm{HSVT},k}$.

Equivalence with PCR. We now state a simple, yet key relation between PCR and the algorithm above. Precisely, the two algorithms produce identical estimated response vectors.

Proposition 2.1. *For any $k \leq N$, $\widehat{Y}^{\mathrm{PCR},k} = \widehat{Y}^{\mathrm{HSVT},k}$.*

By establishing the equivalence above, it allows us to analyze PCR through the growing matrix estimation/completion literature, of which HSVT is one of the most commonly analyzed methods. In fact, there is significant literature establishing that HSVT is a noise-model-agnostic method that recovers the ground-truth matrix given a sparse, noisy observation of it, e.g., see [18]

$\|\cdot\|_{2,\infty}$**-norm error bound for HSVT.** The limitation of the current results concerning HSVT is that they only establish its estimation accuracy in terms of the mean-squared error or expected squared Frobenius norm of the error matrix. To establish our above mentioned results on the prediction error of PCR, it seems necessary to bound the expected squared $\ell_{2,\infty}$-norm of the error matrix (see Lemmas 3.1 and 3.2), which is a stronger guarantee than the Frobenius norm. To see this, let $\boldsymbol{E} = [e_{ij}] \in \mathbb{R}^{n\times p}$ denote the error matrix; then,

$$\frac{1}{np}\|\boldsymbol{E}\|_F^2 = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p} e_{ij}^2 \leq \frac{1}{n}\max_{j\in[p]}\sum_{i=1}^{n} e_{ij}^2 = \frac{1}{n}\|\boldsymbol{E}\|_{2,\infty}^2.$$

Given the ubiquity of HSVT, the $\|\cdot\|_{2,\infty}$-norm result for HSVT may be of interest in its own right.

2.2 Connecting PCR to Synthetic Controls

We briefly describe the application of the analysis of PCR to SC, which has become a standard method in econometrics (and beyond) to make counterfactual predictions utilizing only observational data.

Robust synthetic control. In [5], the authors propose the robust synthetic control (RSC) method, which pre-processes observations using HSVT before performing linear regression to learn the model. They observed empirically that the resulting synthetic control had attractive robustness properties such as robustness to noisy and partially observed data, and thus suggested an alternative model to the convex weights originally proposed by [2, 1]; compare Figures 4b and 6c with Figures 2b and 6b. Using Proposition 2.1, we establish PCR is identical to the RSC estimator. This provides empirical evidence of the importance of pre-processing the covariates (in the setting of SC, this is the donor pool data) by finding its low-dimensional representation. See Section 4 for details.

3 Main Results

Notations. For any matrix $\boldsymbol{B} \in \mathbb{R}^{N \times p}$, let $\|\boldsymbol{B}\|_F, \|\boldsymbol{B}\|_2, \|\boldsymbol{B}\|_\infty$ denote the Frobenius norm, operator norm, and max norm (i.e., largest absolute value among all entries) of a matrix $\boldsymbol{B}$, respectively; let $\|\boldsymbol{B}\|_{2,\infty}$ denote the max column ℓ_2-norm of $\boldsymbol{B}$. For an index set $\Omega \subset [N]$, let $\boldsymbol{B}^\Omega$ denote the $|\Omega| \times p$ submatrix of $\boldsymbol{B}$ formed by stacking the rows of $\boldsymbol{B}$ according to Ω, i.e., $\boldsymbol{B}^\Omega$ is the concatenation of $\{\boldsymbol{B}_{i,\cdot} : i \in \Omega\}$. The superscript Ω is sometimes omitted if the matrix representation is clear from context. Let $x \vee y = \max(x,y)$ and $x \wedge y = \min(x,y)$ for any $x,y \in \mathbb{R}$. Lastly, let $\mathbb{1}$ denote the indicator function.

3.1 Key Modeling Assumptions

We recall the approximate linear model given by (1).

Bounded covariates. We assume the entries of $\boldsymbol{A}$ are bounded. Without loss of generality, we assume the entries are bounded by 1.

Property 3.1. *The entries of $\boldsymbol{A}$ are bounded by one in absolute value, i.e., $\|\boldsymbol{A}\|_\infty \leq 1$.*

Noise on response variables. We make the standard assumption that the noise on the response variable, denoted by ϵ_i, is mean zero and has bounded variance.

Property 3.2. *The response noise $\epsilon = [\epsilon_i] \in \mathbb{R}^N$ is a random vector with independent, mean zero entries such that each of its components has variance bounded above by σ^2.*

Noise on covariates. Recall that rather than observing $\boldsymbol{A}$, we are given access to its partially observed and noisy version $\boldsymbol{Z}$. Let $\boldsymbol{H} = [\eta_{ij}] \in \mathbb{R}^{N \times p}$ denote the covariate noise matrix. We define $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{H}$ as the noisy perturbation of the covariate matrix, without missing values. We assume the following property about the noise matrix $\boldsymbol{H}$ (see Definition E.1 for the definition of ψ_α-random variables/vectors).

Property 3.3. *Let $\boldsymbol{H}$ be a matrix of independent, mean zero ψ_α-rows for some $\alpha \geq 1$, i.e., there exists an $\alpha \geq 1$ and $K_\alpha < \infty$ such that $\|\eta_{i,\cdot}\|_{\psi_\alpha} \leq K_\alpha$ for all $i \in [N]$. Further, assume there exists a $\gamma^2 > 0$ such that $\|\mathbb{E}\eta_{i,\cdot}^T \eta_{i,\cdot}\|_2 \leq \gamma^2$ for all $i \in [N]$. Lastly, for all $i \in [N], j \in [p]$, assume variance of $\eta_{i,j}$ is bounded above σ^2.*

Remark 3.1. *One can verify that if the entries of $\eta_{i,\cdot}$ are independent, then $\gamma^2 = \mathcal{O}(1)$ (e.g., for independent standard normal random variables, $\gamma^2 = 1$). In general, γ^2 will scale linearly with the number of correlated entries in $\eta_{i,\cdot}$; similarly, K_α scales with the square root of the correlated entries.*

Remark 3.2. *We assume that the response noise ϵ and covariate noise $\boldsymbol{H}$ are independent of each other. Further, if we denote $D_{ij} \in \{0,1\}$ as the random variable indicating whether Z_{ij} is missing or not, we assume D_{ij} is independent of ϵ and $\boldsymbol{H}$. Relaxing these assumptions and allowing for dependencies between these three sources of noise remains interesting future work.*

3.2 Training Prediction Error

In this section, we present bounds on the training error under different settings.

3.2.1 General Results

We first state Theorem 3.1 (proof in Appendix G), which bounds the training error of PCR in terms of three natural quantities, as described below.

Theorem 3.1 (Training Error of PCR: Generic Result)**.** *Consider PCR with parameter $k \geq 1$. Suppose Property 3.2 holds. Then, under the model described by* (1),

$$\mathrm{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 k}{n} + \frac{3\|\beta^*\|_1^2}{n}\mathbb{E}\|(\boldsymbol{Z}^{\mathrm{HSVT},k,\Omega} - \boldsymbol{A}^\Omega)\|_{2,\infty}^2 + \frac{20\|\phi\|_2^2}{n} \tag{4}$$

Interpretation. The bound in (4) has three terms on the right hand side: (a) $\sigma^2 k/n$ represents the standard "regression" prediction error, which scales with the model complexity k and inversely with number of samples n; (b) $(1/n)\|\beta^*\|_1^2\mathbb{E}\|\boldsymbol{Z}^{\mathrm{HSVT},k,\Omega} - \boldsymbol{A}^\Omega\|_{2,\infty}^2$, which is a consequence of the corruption of $\boldsymbol{A}$ (if $\boldsymbol{A}$ was fully observed and rank k, then this error term would vanish); (c) $(1/n)\|\phi\|_2^2$ represents the (inevitable) impact of the model mismatch.

Quantification. To quantify (4), we need to evaluate $\mathbb{E}[\|\boldsymbol{Z}^{\mathrm{HSVT},k,\Omega} - \boldsymbol{A}^\Omega\|_{2,\infty}^2]$, where $\boldsymbol{Z}^{\mathrm{HSVT},k}$ is the estimate of $\boldsymbol{A}$ produced from the sparse, noisy observation of it, $\boldsymbol{Z}$. Our interest is in evaluating the estimation error with respect to the $\ell_{2,\infty}$-error. As stated earlier, the estimation error for HSVT is typically evaluated with respect to the Frobenius norm and this quantity is well understood, e.g., see [18]. On the other hand, the error bound with respect to $\ell_{2,\infty}$-norm is unknown. To that end, we provide a novel characterization of this error in Lemma 3.1 below (proof in Appendix I).

Let $\boldsymbol{A} = \sum_{i=1}^N \tau_i u_i v_i^T$ with its singular values τ_i arranged in descending order. Let $\boldsymbol{A}^k = \sum_{i=1}^k \tau_i u_i v_i^T$ denote the truncation of $\boldsymbol{A}$ obtained by retaining the top k components.

Lemma 3.1 ($\ell_{2,\infty}$-error bound for HSVT)**.** *Let Properties 3.1, 3.2, 3.3 hold. If $\rho \geq 64\log(Np)/(Np)$, then*

$$\mathbb{E}[\|\boldsymbol{Z}^{\mathrm{HSVT},k} - \boldsymbol{A}\|_{2,\infty}^2] \leq \frac{C'}{\rho^4}\left(\frac{N(N \vee p)}{(\tau_k - \tau_{k+1})^2} + k\right)\log^5(Np) + 2\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2,$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ and $C > 0$ is an absolute constant.

3.2.2 Low-Rank Covariates, Well-Balanced Spectra

We state the following result for PCR when the covariate matrix is low-rank, i.e., $\boldsymbol{A}$ admits a low-dimensional representation, and PCR chooses the correct number of principal components.

Property 3.4. *Let r denote the rank of $\boldsymbol{A}$. The r-th largest singular value (i.e., the smallest nonzero singular value) of A satisfies $\tau_r = \Omega(\sqrt{Np/r})$.*

Property 3.4 combined with Property 3.1 imply the singular spectrum of $\boldsymbol{A}$ is "well-balanced" in the sense that $\frac{\tau_1}{\tau_r} = O(\sqrt{r})$. Below, we describe another natural setting under which Property 3.4 holds.

Remark 3.3. *A natural setting in which Property 3.4 holds is if $\boldsymbol{A} = \Theta(1)$ and the non-zero singular values of $\boldsymbol{A}$ satisfy $\tau_i^2 = \Theta(\zeta)$ for some ζ. Then, $Cr\zeta = \|\boldsymbol{A}\|_F^2 = \Theta(Np)$ for some constant C, i.e., $\tau_i^2 = \Theta(Np/r)$. See Proposition 3.1 below for a canonical probabilistic generating process used to analyze probabilistic PCA in [14, 40], under which Property 3.4 holds.*

Corollary 3.1. *Let Properties 3.1, 3.2, 3.3, 3.4 hold. Suppose PCR chooses the correct number of principal components $k = r = \text{rank}(\boldsymbol{A})$. Let $\rho \geq 64 \log(Np)/(Np)$ and $n = \Theta(N)$. Then for any given $\Omega \subset [N]$,*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 r}{n} + \frac{C'\|\beta^*\|_1^2}{\rho^4} \frac{r \log^5(np)}{n \wedge p} + \frac{20\|\phi\|_2^2}{n}, \tag{5}$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ and $C > 0$ is an absolute constant.

Proof. Corollary 3.1 follows from Theorem 3.1 and Lemma 3.1 by setting $k = r$, $\boldsymbol{A}^k = \boldsymbol{A}$, $\tau_{k+1} = 0$. □

Interpretation. The statement of Corollary 3.1 requires that the *correct* number of principal components are chosen in PCR. In settings where all r singular values of $\boldsymbol{A}$ are roughly equal (Property 3.4), the training prediction decays (up to logarithmic factors) as $\rho^{-4} r/(n \wedge p) + \|\phi\|_2^2/n$. We note that for this exact low-rank setting, analyzing the case where $k > r$ (e.g., as done in [33]) is interesting future work.

Example: embedded Gaussian features. We present a classical data generating process under which PCR (and PCA) is justified. Consider the setting where $\boldsymbol{A} \in \mathbb{R}^{N\times p}$ is generated by sampling its rows from a distribution on $\mathbb{R}^p$, which in turn, is an embedding of some underlying latent distribution on $\mathbb{R}^r$; this is similar in spirit to the probabilistic model for PCA, cf. [14, 40].

Proposition 3.1. *Let $\boldsymbol{A} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{R}}$, where the entries of $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N\times r}$ are independent standard normal random variables, i.e., $\tilde{A}_{ij} \sim \mathcal{N}(0,1)$ and $\tilde{\boldsymbol{R}} \in \mathbb{R}^{r\times p}$ is another random matrix with independent entries drawn uniformly at random from $\{-1/\sqrt{r}, 1/\sqrt{r}\}$. Suppose, $r \leq \frac{\sqrt{p}}{4\sqrt{2\log p}} + 1$ and $r = o(N)$ Then $\|\boldsymbol{A}\|_\infty \leq 4\sqrt{\log(Np)}$ and Property 3.4 holds with probability at least $1 - \frac{2}{N^2 p} - 2\exp\left(-c\sqrt{Nr}\right)$ for some constant $c > 0$.*

In a strict sense, $\boldsymbol{A}$ in Proposition 3.1 does not satisfy Property 3.1 because of the extra log factor. Taking a closer look at the proof of Lemma 3.1, we can see that this slack only makes the exponent of the log slightly larger ($5 \rightarrow 6$) in Lemma 3.1 and Corollary 3.1. Proof of Proposition 3.1 can be found in Appendix N.1.

3.2.3 Beyond Low-Rank Covariates—Low-Rank Approximation in $\|\cdot\|_2$-norm

In Corollary 3.2, we generalize the result of Corollary 3.1 to the setting where the low-rank model is misspecified, i.e., $\boldsymbol{A}$ does not equal $\boldsymbol{A}^k$.

Corollary 3.2. *Let Properties 3.1, 3.2, 3.3 hold. Suppose $\rho \geq 64\log(Np)/(Np)$. Let $n = \Theta(N)$. Then,*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 k}{n} + \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(\frac{n \vee p}{(\tau_k - \tau_{k+1})^2} + \frac{k}{n}\right)\log^5(np) + \frac{3\|\beta^*\|_1^2}{n}\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2 + \frac{20}{n}\|\phi\|_2^2, \tag{6}$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ and $C > 0$ is an absolute constant.

Proof. Corollary 3.2 follows immediately from Theorem 3.1 and Lemma 3.1. □

Interpretation. Corollary 3.2 implies the training prediction error, not including the linear model mismatch ϕ, decays to zero if: (i) the gap between the k-th and $(k+1)$-st singular values of $\boldsymbol{A}$ grows faster than $n \vee p$

(ignoring log factors); (ii) $k = o(n)$; (iii) $\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2 = o(n)$. Below, we show that if the spectrum of $\boldsymbol{A}$ is geometrically decaying, then there exists a range of k such that Properties 3.1, 3.2, and 3.3 are satisfied.

Example: geometrically decaying singular values. To explain the utility of Corollary 3.2, we consider a setting where $\boldsymbol{A}$ has geometrically decaying singular values, and is thus *approximately* low-rank. We note that such a setting is representative of many real-world datasets; as an example, see Figures 3a and 5a. Further, matrices with geometrically decaying singular values are also ubiquitous models in the study of a variety of domains including graphon estimation and signal processing.

Let $e_{\cdot,j} \in \mathbb{R}^p$ denote the j-th canonical basis vector. Recall that u_i, v_i, and τ_i denote the left singular vectors, right singular vectors, and singular values of $\boldsymbol{A}$, respectively.

Proposition 3.2. *Let Properties 3.1, 3.2, 3.3 hold. Suppose* $\rho \geq 64 \log(Np)/(Np)$. *Let* $n = \Theta(N)$. *Let* $\tau_1 = C_1\sqrt{Np}$ *and* $\tau_k = \tau_1 \theta^{k-1}$ *for all* $k \in [N]$ *with* $\theta \in (0,1)$. *Further, let* $v_i^T e_j = O(1/\sqrt{p})$ *for all* $i,j \in [p]$. *Consider PCR with parameter* $k = \frac{1}{4} \cdot \frac{\log(n \wedge p)}{\log(1/\theta)}$. *Then,*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{C' C(\theta) \|\beta^*\|_1^2}{\rho^4} \frac{\log^6(np)}{(n \wedge p)^{1/2}} + \frac{20}{n}\|\phi\|_2^2, \tag{7}$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$, $C(\theta) > 0$ *depends only on* θ, *and* $C_1 > 0$ *is an absolute constant.*

Proof of Proposition 3.2 can be found in Appendix N.2.

Interpretation. The conditions on the spectrum of $\boldsymbol{A}$ in Proposition 3.2 are self-explanatory with potentially one exception, $v_i^T e_j = O(1/\sqrt{p})$. In effect, this assumption states that the right singular vectors of $\boldsymbol{A}$ satisfy an "incoherence" condition, cf. [15], with the canonical basis of $\mathbb{R}^p$; or, equivalently, all entries of the right singular vectors are roughly of the same magnitude, $O(1/\sqrt{p})$. See Appendix N.3 for an explicit construction of such a matrix from signal processing. The bound in Proposition 3.2 implies that if the number of principal components is chosen as $4(\log(np)/\log(1/\theta))$ and $(n \wedge p) = \Omega(\rho^{-4} \mathsf{poly}(\log p))$, then the training prediction error is dominated by $(1/n)\|\phi\|_2^2$. This is precisely the unavoidable linear model mismatch error.

3.2.4 Beyond Low-Rank Covariates—Low-Rank Approximation in $\|\cdot\|_\infty$-norm

Thus far, in Sections 3.2.2 and 3.2.3, $\boldsymbol{A}$ has been assumed to be well-approximated by a *specific* low-rank matrix $\boldsymbol{A}^k$ that is induced by retaining the top k singular values of $\boldsymbol{A}$. Such an approximation is optimal with respect to Frobenius and spectral norm. However, for approximating with respect to other norms, e.g., $\ell_{2,\infty}$ or ℓ_∞, exciting progress has been made to obtain different styles of low-rank approximations (see for example [43, 48] and references therein). Indeed, such low-rank approximations of $\boldsymbol{A}$ may not correspond to $\boldsymbol{A}^k$. For that reason, we provide an analogous result to Lemma 3.1 and Corollary 3.2 for the setting when $\boldsymbol{A}$ is well-approximated by some *arbritrary* low-rank matrix.

Specifically, let $\boldsymbol{A}=\boldsymbol{A}^{(\mathrm{lr})}+\boldsymbol{E}^{(\mathrm{lr})}$. In words, $\boldsymbol{A}^{(\mathrm{lr})}$ denotes a low-rank matrix and $\boldsymbol{E}^{(\mathrm{lr})}$ denotes the approximation error between $\boldsymbol{A}$ and $\boldsymbol{A}^{(\mathrm{lr})}$. Let $r=\mathrm{rank}(\boldsymbol{A}^{(\mathrm{lr})})$ and let the SVD of $\boldsymbol{A}^{(\mathrm{lr})}=\sum_{i=1}^r \tau_i u_i v_i^T$ (again, with the singular values τ_i arranged in descending order).

Lemma 3.2. *Let Properties 3.1, 3.2, 3.3 hold. Consider PCR with parameter $k=r$. Let $\rho \geq 64\log(Np)/(Np)$. Then,*

$$\mathbb{E}[\|\boldsymbol{Z}^{\mathrm{HSVT},k}-\boldsymbol{A}\|_{2,\infty}^2] \leq \frac{C'}{\rho^4}\left(\frac{N(N\vee p\vee\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2)}{\tau_r^2}+r\right)\log^5(Np)+2\|\boldsymbol{E}^{(\mathrm{lr})}\|_{2,\infty}^2,$$

where $C'=C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ and $C>0$ is an absolute constant.

Proof of Lemma 3.2 can be found in Appendix J.

Interpretation. Lemma 3.2 is similar to the result of Lemma 3.1; however, because we now only assume that $\boldsymbol{A}$ is well-approximated by *an* arbitrary low-rank matrix $\boldsymbol{A}^{(\mathrm{lr})}$ rather than $\boldsymbol{A}^k$ as done in Section 3.2.3, this introduces an additional $\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2/\tau_r^2$ term compared to the bound in Lemma 3.1.

Corollary 3.3. *Let Properties 3.1, 3.2, 3.3 hold. Consider PCR with parameter $k=r$. Let $\rho \geq 64\log(Np)/(Np)$. Let $n=\Theta(N)$. Then,*

$$\mathrm{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 r}{n}+\frac{C'\|\beta^*\|_1^2}{\rho^4}\left(\frac{n\vee p\vee\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{\tau_r^2}+\frac{r}{n}\right)\log^5(np)+\frac{6\|\beta^*\|_1^2}{n}\|\boldsymbol{E}^{(\mathrm{lr})}\|_{2,\infty}^2+\frac{20}{n}\|\phi\|_2^2, \quad (8)$$

where $C' = CK_\alpha^2(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ *and* $C > 0$ *is an absolute constant.*

Proof. Corollary 3.3 follows immediately from Theorem 3.1 and Lemma 3.2. □

Interpretation. If $\boldsymbol{A}^{(\mathrm{lr})}$ satisfies Property 3.4, i.e., the well-balanced spectra condition, then one can verify (again, ignoring log factors) the prediction error scales as $\rho^{-4}r/(n \wedge p) + \|\phi\|_2^2/n + \rho^{-4}r\|\boldsymbol{E}^{(\mathrm{lr})}\|_\infty^2$. This is identical to the bound in Corollary 3.1 with an additional $\rho^{-4}r\|\boldsymbol{E}^{(\mathrm{lr})}\|_\infty^2$ term, which arises since $\boldsymbol{A}$ is not assumed to be low-rank but rather is well-approximated by an arbitrary low-rank matrix $\boldsymbol{A}^{(\mathrm{lr})}$. Below, we show that under a generalized factor model, $r\|\boldsymbol{E}^{(\mathrm{lr})}\|_\infty^2$ vanishes to zero as n grows.

Example: generalized factor model. We say the $\boldsymbol{A}$ is generated as per a generalized factor model or latent variable model (LVM) if

$$A_{ij} = g(\theta_i, \rho_j), \tag{9}$$

where $\theta_i \in \mathbb{R}^{d_1}$ and $\rho_j \in \mathbb{R}^{d_2}$ are latent features that capture measurement i and feature j specific information, respectively, for some $d_1, d_2 \geq 1$; and the latent function $g : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ captures the model relationship. If g is "well-behaved", e.g., Hölder continuous, and the latent spaces are compact, then Proposition 3.3 shows $\boldsymbol{A}$ is well approximated by a low-rank matrix with respect to the $\|\cdot\|_\infty$-norm, where the approximation error vanishes as more data is collected.

Hölder continuous functions. We now define the Hölder class of functions, which is widely adopted in the non-parametric regression literature (see [48, 41]). Given a function $g : [0,1)^K \to \mathbb{R}$, and a multi-index $\kappa \in \mathbb{N}^K$, let the partial derivate of g at $x \in [0,1)^K$ (if it exists) be denoted as

$$\nabla_\kappa g(x) = \frac{\partial^{|\kappa|} g(x)}{(\partial x)^\kappa} = \frac{\partial^{|\kappa|} g(x)}{(\partial x_1)^{\kappa_1} (\partial x_2)^{\kappa_2} \ldots (\partial x_K)^{\kappa_K}}. \tag{10}$$

Definition 3.1 (**($\zeta, \mathcal{L}$)-Hölder Class**)**.** *Let* $\zeta, \mathcal{L}$ *be two positive numbers. The Hölder class* $\mathcal{H}(\zeta, \mathcal{L})$ *on*

$[0,1)^K$ *is defined as the set of functions* $g:[0,1)^K \to \mathbb{R}$ *whose partial derivatives satisfy*

$$\sum_{\kappa:|\kappa|=\lfloor\zeta\rfloor} \frac{1}{\kappa!} |\nabla_\kappa g(x) - \nabla_\kappa g(x')| \leq \mathcal{L} \|x-x'\|_\infty^{\zeta-\lfloor\zeta\rfloor} \quad \textit{for all } x,x' \in [0,1)^K. \tag{11}$$

Here, $\lfloor\zeta\rfloor$ *denotes the largest integer strictly smaller than* ζ*. We note that the domain is easily extended to any compact subset of* $\mathbb{R}^K$*.*

Remark 3.4. *Note if* $\zeta \in (0,1]$*, then* (11) *is equivalent to the* $(\zeta,\mathcal{L})$*-Lipschitz condition, i.e.,*

$$|g(x)-g(x')| \leq \mathcal{L}\|x-x'\|_\infty^{\zeta-\lfloor\zeta\rfloor} \quad \textit{for all } x,x' \in [0,1)^K.$$

However, for $\zeta > 1$*,* $(\zeta,\mathcal{L})$*-Hölder smoothness no longer implies* $(\zeta,\mathcal{L})$*-Lipschitz smoothness.*

Proposition 3.3. *Let* $\boldsymbol{A}$ *satisfy* (9) *with* $\theta_i,\rho_j \in [0,1)^K$ *as latent parameters. Further, for all* ρ_j*, let* $g(\cdot,\rho_j) \in \mathcal{H}(\zeta,\mathcal{L})$ *as defined in* (11)*. Then, for any* $\delta > 0$*, there exists a low-rank matrix* $\boldsymbol{A}^{(\mathrm{lr})}$ *of rank* $r \leq C(\zeta,K)\delta^{-K}$ *such that* $\left\|\boldsymbol{A} - \boldsymbol{A}^{(\mathrm{lr})}\right\|_\infty \leq \mathcal{L}\cdot\delta^\zeta$*. Here,* $C(\zeta,K)$ *is a term that depends only on* ζ *and* K*.*

The proof of Proposition 3.3 can be found in Appendix O.1.

Remark 3.5. *We remark on the Hölder continuity of a typical linear factor model, i.e.,* $g(\theta_i,\rho_j) = \langle\theta_i,\rho_j\rangle$ *for some latent vectors* $\theta_i,\rho_j \in \mathbb{R}^K$*. It is easily seen that such a model satisfies Definition 3.1 for all* $\zeta \in \mathbb{N}$*, and* $\mathcal{L} = C$*, for some absolute positive constant,* C*. Thus, one can think of Hölder continuous functions as generalizations of typical linear factor models to sufficiently smooth non-linear functions.*

Corollary 3.4. *Let Properties 3.1, 3.2, 3.3 hold. Consider PCR with* $k = r$*. Let* $\rho \geq 64\log(Np)/(Np)$*. Let* $n = \Theta(N)$*. Let the conditions of Proposition 3.3 hold and further assume* $\boldsymbol{A}^{(\mathrm{lr})}$ *(as defined in Proposition 3.3) satisfies Property 3.4. Then,*

$$\mathrm{MSE}_\Omega(\widehat{Y}) \leq \frac{C'C(\zeta,K)\mathcal{L}^2\|\beta^*\|_1^2}{\rho^4}\left(\frac{1}{(n\wedge p)^{1-\frac{K}{2\zeta}}}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2, \tag{12}$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ *and* $C > 0$ *is an absolute constant.*

The proof of Corollary 3.4 can be found in Appendix K. Note that as long as $\zeta > K/2$, this leads to vanishing training error.

3.3 Test Prediction Error

We now evaluate the generalization performance of PCR. As previously mentioned, the emphasis of this work is to provide a rigorous analysis on the prediction properties of the PCR algorithm through the lens of HSVT. Recall from Proposition 2.1, PCR with parameter k is equivalent to linear regression with pre-processing of the noisy covariates using HSVT where the top k singular values are retained. To that end, we study candidate vectors $\beta^{\text{HSVT},k} = \boldsymbol{V}_k \cdot \beta^{\text{PCR},k} \in \mathbb{R}^p$. In light of this observation, we establish the following simple but useful result that suggests restricting our model class to sparse linear models only (the proof of which can be found in Appendix L).

Proposition 3.4. *Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\operatorname{rank}(\boldsymbol{X}) = k$. Without loss of generality, let $\{\boldsymbol{X}_{\cdot,1},...,\boldsymbol{X}_{\cdot,k}\}$ form a collection of k linearly independent vectors, i.e., for any $i \in \{k+1,...,p\}$, there exists some $c(i) \in \mathbb{R}^k$ such that $\boldsymbol{X}_{\cdot,i} = \sum_{\ell=1}^{k} c_l(i) \boldsymbol{X}_{\cdot,\ell}$. Assume the following condition on $\boldsymbol{X}$ holds:*

$$\max_{i \in \{k+1,...,p\}} \|c(i)\|_\infty \leq C''. \tag{13}$$

Then if, $M = \boldsymbol{X} v$ for some $v \in \mathbb{R}^p$, there exists $v^* \in \mathbb{R}^p$ such that $M = \boldsymbol{X} v^*$, $\|v^*\|_0 = k$, and $\|v^*\|_1 \leq C'' k \|v\|_1$.

Interpretation. By Proposition 3.4, for any $\boldsymbol{Z}^{\text{HSVT},k}$ and $\beta^{\text{HSVT},k} = \boldsymbol{V}_k \beta^{\text{PCR},k}$, there exists a $\beta' \in \mathbb{R}^p$ such that $\boldsymbol{Z}^{\text{HSVT},k} \beta^{\text{HSVT},k} = \boldsymbol{Z}^{\text{HSVT},k} \beta'$ where $\|\beta'\|_0 \leq k$ and $\|\beta'\|_1 \leq k \|\beta^{\text{HSVT},k}\|_1$. Thus, for the purposes of bounding the test error of PCR with parameter k via the toolkit of Rademacher complexity, *it suffices to restrict our hypothesis class to linear predictors with sparsity k.* Condition (i) in Proposition 3.4 is a mild assumption circumventing the pathological case that the linear coefficients used to represent columns $\boldsymbol{X}_{\cdot,i}$ for $i \in \{k+1,...,p\}$ in terms of $\{\boldsymbol{X}_{\cdot,1},...,\boldsymbol{X}_{\cdot,k}\}$ are unbounded.

Theorem 3.2 (Test Error of PCR)**.** *Let Property 3.1 hold. Let* $n = \Theta(N)$. *Consider PCR with parameter* $k \geq 1$ *and assume* $\boldsymbol{Z}^{\mathrm{HSVT},k}$ *satisfies* (13) *in Proposition 3.4. Then*

$$\mathbb{E}_{\Omega}\Big[\mathrm{MSE}(\widehat{Y})\Big] \leq \mathbb{E}_{\Omega}\Big[\mathrm{MSE}_{\Omega}(\widehat{Y})\Big] + \frac{C''' k^{5/2}}{\sqrt{n}} \|\beta^*\|_1, \tag{14}$$

where $C''' = C \cdot C'' \cdot \mathbb{E}[\|\beta^{\mathrm{HSVT},k}\|_1^2 \cdot \|\boldsymbol{Z}^{\mathrm{HSVT},k}\|_\infty^2]$*, with* C'' *defined as in Proposition 3.4 and* $C > 0$ *is an absolute constant;* $\mathbb{E}_\Omega$ *denotes the expectation taken with respect to* $\Omega \subset [N]$ *(of size* n*), which is chosen uniformly at random without replacement.*

See Appendix M for a proof of Theorem 3.2.

Interpretation. We note all our training error bounds do not depend on Ω; see (4), (21), (6), (7), (8), (12). Hence, the bound on $\mathbb{E}_\Omega[\mathrm{MSE}_\Omega(\widehat{Y})]$ also does not depend on Ω for these settings. Note the test error decays at a rate $1/\sqrt{n}$, in comparison with $1/n$ for the training error. This "slow rate" of $1/\sqrt{n}$ for test error is indeed the best achievable using the standard Rademacher complexity analysis (see Chapter 4 of [46]). An important open problem is to achieve the fast rate of $1/n$ for the test error in the error-in-variables setting; we remark that a related work [4] takes key steps towards solving this.

Choosing k. We describe how the test prediction error can help in choosing the parameter for PCR (model complexity) in a data-driven manner. Specifically, Theorem 3.2 suggests that the overall error is at most the training error plus a term that scales as $k^{5/2}/\sqrt{n}$. Therefore, one should choose the k that minimizes this bound. Naturally, as k increases, the training error is likely to decrease, but the additional term $k^{5/2}/\sqrt{n}$ will increase; an optimal k can thus be found in a data-driven manner.

3.4 Discussion

Comparison with ordinary least squares (OLS). It is known that OLS implicitly performs regularization if the covariates $\boldsymbol{A}$ are exactly low-rank, noiseless, and fully observed (see Lemma 3.1 of [36]). In most real-world settings, however, data is never precisely low-rank, but is rather *approximately* low-rank, such as in the examples detailed in Sections 3.2.3 and 3.2.4 (see [42] and references therein for further theoretical justification for approximately low-rank covariate matrices). In such a setting, it is not established, nor is it likely, that OLS has the same implicit regularization effect as before. Indeed, in the example shown in Figure 2c, OLS has very poor empirical generalization performance even though over 99% of the spectral energy is captured in the top singular value, i.e., the covariate matrix is very-well approximated by a rank-one matrix. In contrast, if the principal components are chosen correctly, then PCR continues to have the desired regularization property, even in the approximate low-rank case. The contrast can be seen in Figure 4a. Additionally, we provide the explicit tradeoff between training and testing error based on the number of selected principal components k.

"Information" spread across covariates is necessary. Within the high-dimensional (error-in-variables) regression literature, there are several different structural assumptions required of the covariate matrix to achieve vanishing prediction or parameter estimation error (see [34] and references therein for some detailed examples). Intuitively, these assumptions state that the signal is "well-spread" across the various columns of the covariate matrix. Below, we consider a simple yet illustrative example for which both PCR and traditional methods from the literature do not seem to provide meaningful answers.

Example. Suppose $\boldsymbol{A}_{\cdot,1} = e_1$ and $\boldsymbol{A}_{\cdot,2} = e_2$, where e_1, e_2 are the canonical basis vectors in $\mathbb{R}^N$, and $\boldsymbol{A} = [\boldsymbol{A}_{\cdot,1}, \boldsymbol{A}_{\cdot,2}, \ldots, \boldsymbol{A}_{\cdot,2}] \in \mathbb{R}^{N \times p}$. Then, it is clear that $r = \text{rank}(\boldsymbol{A}) = 2$.

What happens to PCR. To estimate $\boldsymbol{A}$, even with the additional (oracle) knowledge of the positions of $\boldsymbol{A}_{\cdot,1}$ and $\boldsymbol{A}_{\cdot,2}$, one can verify the optimal estimators for $\boldsymbol{A}_{\cdot,1}$ and $\boldsymbol{A}_{\cdot,2}$ are $\boldsymbol{Z}_{\cdot,1}$ and $1/(p-1)\sum_{j=2}^{p} \boldsymbol{Z}_{\cdot,j}$, respectively. This results in the following lower bound on the recovery error

$$\|\widehat{\boldsymbol{A}} - \boldsymbol{A}\|_{2,\infty}^2 \geq \|\boldsymbol{Z}_{\cdot,1} - \boldsymbol{A}_{\cdot,1}\|_2^2 = \|\eta_{\cdot,1}\|_2^2 \overset{\mathbb{E}}{=} N,$$

yielding

$$\frac{1}{N}\mathbb{E}\left[\|\widehat{Y} - \boldsymbol{A}\beta^*\|_2^2\right] \leq \frac{\sigma^2 r}{N} + \|\beta^*\|_1,$$

which does not lead to prediction consistency. In fact, the second term, $\|\beta^*\|_1$, is exactly what arises if the bias is not corrected in the error-in-variables regression setting of [31, 37, 11, 12].

What happens to traditional error-in-variables regression estimators. Now, consider the same setup as above but let $\boldsymbol{A}$ be fully observed, i.e., $\boldsymbol{Z}=\boldsymbol{A}$. In such settings where $\boldsymbol{A}$ is uncontaminated, it is known that the restricted eigenvalue (RE) condition (see Definition E.2 of Appendix E), which is the de-facto assumption in the literature, guarantees ℓ_2-recovery of the underlying β^* via the Lasso method. However, this particular $\boldsymbol{A}$ breaks the RE condition and thus β^* cannot be accurately estimated. To see this, let $\Delta=e_3\in\mathbb{R}^p$ in Definition E.2. Then, $(1/N)\|\boldsymbol{A}\Delta\|_2^2=0$, hence violating the RE condition needed for all the existing analyses of methods for the error-in-variables regression setting.

In summary. From this simple example, we observe that a lack of information spread across the columns of $\boldsymbol{A}$ seem to yield poor prediction and parameter estimation errors. However, it has been well established that a large ensemble of covariate matrices $\boldsymbol{A}$ satisfy the RE condition; specifically, when the entries (or rows) of $\boldsymbol{A}$ are sampled independently from a sub-gaussian distribution. Analogously, we show that two canonical generating processes for the covariate matrix, namely embedded Gaussian features and geometrically decaying singular values, satisfy the desired properties needed to achieve vanishing prediction error. That is, the singular value gap $\tau_k-\tau_{k+1}$ is sufficiently large under such generating processes for appropriate k, where k is the number of chosen principal components.

4 PCR and Synthetic Controls

4.1 Synthetic Controls Setup

Pre- & post-intervention periods. As is standard in the SC literature, let there be $p+1$ different time series over N periods associated with a target unit and p donor units. Suppose the target unit receives the intervention at time period n, where $1\leq n<N$. We will refer to the pre- and post- intervention

periods as the time periods prior to and after the intervention point.

Donor observations under control. Let $\boldsymbol{A} \in \mathbb{R}^{N \times p}$ represent the true utilities of the p donor units across the entire time horizon N in the absence of intervention; i.e., $\boldsymbol{A}_{\cdot,j} \in \mathbb{R}^{N}$ represents the time series over N periods for donor $j \in [p]$. Rather than observing $\boldsymbol{A}$, we assume we are only given access to $\boldsymbol{Z} \in \mathbb{R}^{N \times p}$, a sparse, noisy instantiation of $\boldsymbol{A}$. In words, $\boldsymbol{Z}$ denotes the corrupted donor pool observations; as made precise later in the section, we assume $\boldsymbol{Z}$ follows the distributional characteristics described in Section 3.1.

Target unit observations under control. For every $i \in [N]$, let Y_i denote the noisy utility associated with the target unit in the absence of intervention (control). However, since the target unit experiences an intervention for all time instances $n < i \leq N$, we only have access to a noisy version of the target unit's utility for the pre-intervention period, i.e., we only observe $Y^{\text{pre}} = [Y_i]$ for $i \in [n]$. Analogously, we denote $Y^{\text{post}} = [Y_i]$ for $i \in N \setminus [n]$ as the target's (noisy) utility in the post-intervention period. We will denote $\mathbb{E}[Y_i] \in \mathbb{R}$ as the true, latent utility at time i for the target unit, if the intervention never occurred. In summary, given data $(Y^{\text{pre}}, \boldsymbol{Z})$, the aim is to recover $\mathbb{E}[Y^{\text{post}}]$, the counterfactual trajectory of the target unit under control in the post-intervention period. For a pictorial view of the setup of the problem, please refer to Figure 1.

4.2 (Approximate) Linear Synthetic Controls Exist

Existence of linear synthetic controls. In the SC literature, two standard assumptions are made: first, there exists a linear relationship between the target and donor units—in [1, 2], a more restrictive assumption is made that a convex relationship between the target and units exists; second, the underlying utilities follow a low-rank factor model.

Below, we show that if the underlying utilities of the target and donor units follow a generalized factor model or latent variable model (LVM) as in (9), then an (approximate) linear relationship between the target and donor units is actually *implied* by such a model. That is, the existence of an (approximate) linear synthetic control does not need to be additionally assumed. Further, we establish that the linear approximation error goes to zero the more data that is collected. As stated in Section 3.2.4, LVMs are a natural nonlinear generalization of the typical *factor model* ubiquitous in studying panel data in econometrics.

To that end, let $\boldsymbol{A'} = [A'_{ij}] \in \mathbb{R}^{N \times (p+1)}$ denote the concatenation of $\boldsymbol{A}$, the latent donor pool utilities, with $\mathbb{E}[Y]$, the vector of underlying utilities for the target unit in the absence of an intervention. We denote $\boldsymbol{A'}_{\cdot,0} = \mathbb{E}[Y]$ as the latent true utility vector for the target unit, and $\boldsymbol{A'}_{\cdot,j} = \boldsymbol{A}_{\cdot,j}$ for all $j \in [p]$ as the latent true utilities for the donor pool. We assume $\boldsymbol{A'}$ follows a LVM as detailed below:

Property 4.1. *Let $\boldsymbol{A'}$ follow a LVM as defined in* (9) *– as established in Proposition 3.3, for any $\delta > 0$, there exists $\boldsymbol{A'}^{(\mathrm{lr})}$ of rank $r \le C(\zeta, K)\delta^{-K}$ such that $\left\| \boldsymbol{A'} - \boldsymbol{A'}^{(\mathrm{lr})} \right\|_{\infty} \le \boldsymbol{\mathcal{L}} \cdot \delta^{\zeta}$. Let $\zeta > K$. Denote $\boldsymbol{A'}^{(\mathrm{lr})} = \boldsymbol{U}\boldsymbol{V}^T$ as its singular value decomposition where $\boldsymbol{U} \in \mathbb{R}^{N \times r}, \boldsymbol{V} \in \mathbb{R}^{(p+1) \times r}$ and v_i denotes the i-th row of $\boldsymbol{V}$. Let v_0 lie within span$(\{v_i\}_{i \in [p]})$.*

Interpretation. If $\boldsymbol{A'}$ satisfies a LVM as defined in (9), then the existence of $\boldsymbol{A'}^{(\mathrm{lr})}$ as in Property 4.1 is simply a restatement of Proposition 3.3. To analyze SC, we make the additional mild assumption in Property 4.1 that v_0 lies within span$(\{v_i\}_{i \in [p]})$. This assumption helps avoid the "pathological" case where the right singular vector associated with the target unit, v_0, does not lie within the span of the right singular vectors associated with the donor units, $\{v_i\}_{i \in [p]}$. Even in the worst case, by the definition of a rank of a matrix, there can only exist r out of the p right singular vectors v_i that do not lie within the span of the remaining singular vectors. Indeed, we can pick δ as defined in Property 4.1, such that $r = \mathrm{rank}(\boldsymbol{A'}^{(\mathrm{lr})}) \le C(\zeta, K)\delta^{-K} = o(p)$, rendering this pathological case overwhelmingly unlikely to hold. Lastly, we note this assumption that v_0 lies within span$(\{v_i\}_{i \in [p]})$ is implicitly always made in the SC literature.

Proposition 4.1. *Assume $\boldsymbol{A'}$ satisfies Property 4.1. Then there exists a $\beta^* \in \mathbb{R}^p$ such that the target unit (represented by index* 0*) satisfies for all $i \in [N]$, and for any $\delta > 0$,*

$$|A'_{i0} - \sum_{k=1}^{p} \beta^*_k \cdot A'_{ik}| \le C(\zeta, K) \cdot \boldsymbol{\mathcal{L}} \cdot \delta^{(\zeta - K)}. \tag{15}$$

Here, $C(\zeta,K)$ is defined as in Property 4.1.

Proof of Proposition 4.1 can be found in Appendix O.2.

Interpretation. Proposition 4.1 shows that if $\boldsymbol{A}'$ follows a LVM as in (9), then a (approximate) linear synthetic control exists, where the linear misspecification error decays to zero for appropriate choice of δ in (15). Moreover, empirically a LVM is well-motivated – across many real-world datasets, including the canonical SC case studies of California Proposition 99 and terrorism in Basque Country of [1, 2], we see they exhibit an approximate (very) low-rank structure (see Figures 3a, 3b, 5a, and 5b).

4.3 Synthetic Controls and Error-in-variables Regression

SC framework fits error-in-variables regression with model mismatch. If $\boldsymbol{A}'$ satisfies Property 4.1, Proposition 4.1 establishes that we can express the underlying utility of the target unit under no intervention for all $i \in [N]$ as

$$\mathbb{E}[Y_i] = \boldsymbol{A}'_{i0} = \boldsymbol{A}_{i,\cdot}\beta^* + \phi_i. \tag{16}$$

Here $\beta^* \in \mathbb{R}^p$ is defined as in Proposition 4.1, and ϕ_i is the model mismatch bounded by $C(\zeta,K)\cdot\boldsymbol{\mathcal{L}}\cdot\delta^{(\zeta-K)}$. That is, in the SC framework, (1) holds under Property 4.1. In summary Proposition 4.1 reduces the question of interest in SC of estimating $\mathbb{E}[Y^{\text{post}}]$ – the counterfactual trajectory of the target unit under no intervention in the post-intervention period – to that of linear regression with model mismatch. We note that we are in the error-in-variable setting as instead of observing $(\mathbb{E}[Y_i^{\text{pre}}],\boldsymbol{A})$, we only get to observe $(Y^{\text{pre}},\boldsymbol{Z})$.

Restating objective in SC framework. Given (16), we can write the *pre-intervention error* as

$$\text{MSE}_{\text{pre}}(\widehat{Y}) = \frac{1}{n}\mathbb{E}\left[\sum_{i\in[n]}(\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*)^2\right], \tag{17}$$

and the *post-intervention error* as

$$\text{MSE}_{\text{post}}(\widehat{Y}) = \frac{1}{N-n}\mathbb{E}\left[\sum_{i\in[N]\setminus[n]}(\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*)^2\right]. \tag{18}$$

(17) is precisely the training error defined in (2) and (18) is a slightly modified form of (3) since the objective now is to accurately estimate the counterfactual in the absence of any intervention only during the post-intervention stage. Observe that the objective in SC exactly fits the setting of transductive semi-supervised learning as described in Section 1.1.

4.4 Finite-sample Analysis of RSC via PCR

RSC is equivalent to PCR. The RSC method proposed by [5] has exhibited empirical success in duplicating the celebrated results of [1, 2] for the California Proposition 99 and terrorism in Basque Country case studies, respectively, using: (i) only the outcome data, i.e., without any usage of auxiliary covariates (ii) in the presence of noisy data. Under these two conditions, the classical SC algorithm of [1, 2] provides poor post-intervention predictions (see Figures 2b and 6b).

The RSC method is a three step procedure: (i) perform HSVT on the donor matrix (include both pre- and post-intervention data); (ii) linearly regress thresholded donor matrix with pre-intervention data of the target unit to learn linear weights for each of the donors; (iii) apply these linear weights on the post-intervention donor data to estimate the counterfactual trajectory for the target unit.

Observe that the RSC method is precisely the algorithm detailed in Section 2.1, of doing HSVT followed by OLS. Empirically, [5] demonstrated that the RSC method's first step of pre-processing via HSVT effectively de-noises and imputes missing values in the donor observations, which is crucial in building a robust linear synthetic control that has good post-intervention performance. Pleasingly, by Proposition 2.1, we can equivalently interpret the RSC method as simply PCR.

It is worth noting that one of the primary motivations for utilizing convex regression (as proposed in [1, 2]) was to impose sparsity in the number of donors chosen, i.e., enforcing most of the coefficients of the synthetic control to be zero. Rather than introducing sparsity in the original donor space, PCR can be interpreted as introducing sparsity in the subspace induced by the right singular vectors corresponding to the donors since only the top few right singular components are retained. Indeed, as made precise by Proposition 3.4, PCR performs implicit ℓ_0-regularization on the learnt linear model.

Theoretical results. By viewing RSC via the lens of PCR, it allows us to bound the post-intervention prediction error for the target unit. We recall some necessary notation. Recall the definition of $\boldsymbol{A}$ and $\boldsymbol{Z}$ from Section 4.1; the definition of $\boldsymbol{A}'$ from Section 4.2; and denote $\boldsymbol{Z}^{\text{HSVT},k}$ and $\beta^{\text{HSVT},k}$ as the de-noised donor matrix and the fitted linear model outputted from RSC, respectively.

Theorem 4.1. *Let $\boldsymbol{A}$,$\boldsymbol{Z}$ satisfy Properties 3.1, 3.2, 3.3. Let $\boldsymbol{A}'$ satisfy: (i)* (9) *and further assume θ_i for $i \in [N]$, the latent parameters associated with time, are sampled i.i.d from some latent distribution Θ; (ii) Property 4.1 and further assume $\boldsymbol{A}^{(\text{lr})} \in \mathbb{R}^{N\times p}$, the restriction of $\boldsymbol{A}'^{(\text{lr})}$ to the donor units, satisfies Property 3.4. Let $\boldsymbol{Z}^{\text{HSVT},k}$ satisfies* (13) *in Proposition 3.4. Let $N-n=\Theta(n)$. Then,*

$$\text{MSE}_{\text{post}}(\widehat{Y}) \leq \frac{C'C(\zeta,K)\mathcal{L}^2\|\beta^*\|_1^2}{\rho^4}\left(\frac{1}{(n\wedge p)^{1-\frac{K}{2\zeta}}}\right)\log^5(np) + \frac{C'''k^{5/2}}{\sqrt{n}}\|\beta^*\|_1, \tag{19}$$

where: $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$; $C(\zeta,K)$ and $\mathcal{L}$ are defined as in Property 4.1; $C''' = C \cdot C'' \cdot \|\beta^{\text{HSVT},k}\|_1 \cdot \mathbb{E}[\|\boldsymbol{Z}^{\text{HSVT},k}\|_\infty^2]$, with C'' defined as in Proposition 3.4; $C > 0$ is an absolute constant.
The proof of Theorem 4.1 can be found in Appendix O.3.

Interpretation. We highlight that Theorem 3.2 bounds $\mathbb{E}_\Omega[\text{MSE}(\widehat{Y})]$, while Theorem 4.1 bounds $\text{MSE}_{\text{post}}(\widehat{Y})$. That is, Theorem 4.1 differs from Theorem 3.2 in that the set of observations for which we see labels (i.e., observations of the target in the pre-intervention period) is *not assumed to be drawn uniformly at random.* Such an assumption obviously cannot hold in the setting of SC as the pre-intervention period chronologically occurs before the post-intervention period. Instead, we make a more standard assumption that the latent features θ_i, which correspond to different time periods, are sampled i.i.d. from some unknown distribution, Θ. Lastly, we leave it as open problem of how to achieve confidence intervals for the post-intervention error for RSC; one could possibly do so by extending our results to hold in high-probability rather than in expectation.

Comparison with related works in SC. The most relevant results to compare against are Corollary 4.1 (pre-intervention prediction error) and Theorem 4.6 (post-intervention prediction error) in [5]. To begin with, as in standard in the SC literature, [5] does not establish the existence of a synthetic control, rather it simply assumes one exists. Corollary 4.1 in [5] does not show consistency of the RSC method with respect to pre-intervention error as there is an irreducible term, σ^2, the measurement noise in the donor pool, that does not vanish. Further, with respect to post-intervention error, Theorem 4.6 of [5] suffers from the same irreducible σ^2 term. In addition, the authors do not show that the second term of their bound decays to zero as more data is collected. As importantly, in both bounds, it is assumed that the RSC method picks the correct number of singular components, i.e., the rank of the underlying matrix of donor utilities is correctly chosen and is of a lower order compared to the ambient dimensions. In contrast, in our setting, we allow the low-rank condition of the underlying donor matrix to be misspecified, i.e., follows a generalization factor model. Finally, their result does not provide guidance for picking the right parameter k for rank (or in PCR) as done by our result through the generalization or post-intervention error analysis.

Additionally, [6] considers a similar setting where the observed covariates $\boldsymbol{Z}$ are a corrupted version (additive noise model) of the true, underlying covariates $\boldsymbol{A}$, which follow an approximately low-rank factor model, i.e., $\boldsymbol{A}$ cannot have too many large singular values (specifically, refer to Assumption 3 of [6]). However, they do not allow for missing data within $\boldsymbol{Z}$. Here, the authors perform convex regression (with ℓ_2-norm constraints) along both the unit and time axes (unlike standard SC methods, such as RSC, which only consider regression along the unit axis) to estimate the causal average treatment effect. As is classically done in the SC literature, the authors of [6] assume that convex weights exist amongst the rows of $\boldsymbol{A}$; in contrast, we show that (approximate) linear weights are directly implied by a (approximate) low-rank factor model. For this setting, they establish a rigorous asymptotic normality result for their causal estimand of interest, which is the average treatment effect of all treated units over the entire post-intervention period; in contrast, our target causal estimand is the entire post-intervention vector for each treated unit, for which we show mean squared error consistency at rate $1/\sqrt{n}$. The work of [6] complements our own in terms of clarifying the tradeoff between the assumptions made on $\boldsymbol{A}$ (i.e., the low-rank approximation error), the constraints on the synthetic control weights (linear vs. convex), and the target causal estimand. Building on these works to explicitly define the tradeoff between what can be assumed on the spectra of $\boldsymbol{A}$, the synthetic control weights, and the subsequent results one can get for various target causal estimand is an interesting future research direction.

Another work that is less related, but is worth commenting on, as it also heavily relies on matrix estimation techniques for SC, is [7]. Here, the authors consider an underlying low-rank matrix of N units and T measurements per unit, and the entries of the observed matrix are considered "missing" once that unit has been exposed to a treatment. To estimate the counterfactuals, [7] applies a nuclear norm regularized matrix estimation procedure. Some key points of difference are that their performance bounds are with respect to the Frobenius norm over all entries (i.e., units and measurements) in the matrix; meanwhile, we provide a stronger bound that is specific to the single treated unit and only during the post-intervention period. Additionally, the bound of [7] depends on a parameter, which they denote as p_c, that represents the minimum probability of observe all T measurements associated with a given unit. The authors establish consistency of their estimator provided that $p_c \gg 1/\sqrt{T}$. When data is randomly missing, even if the probability of observing each entry is $1-\varepsilon$ for any $\varepsilon > 0$, then $p_c < (1-\varepsilon)^T = o(1/\sqrt{T})$; thus, this result is not applicable for our setting.

4.5 Empirical Results

We present empirical results using the RSC method on several well-known datasets in the literature to highlight its robustness properties in comparison with the traditional SC estimator and OLS.

Terrorism in Basque Country. A canonical case study within the SC literature investigates the impact of terrorism on the economy in Basque Country (see [2]). Here, the target unit of interest is Basque Country, the donor pool consists of neighboring Spanish regions, and the intervention is represented by the first wave of terrorist activity in 1970. The aim in this study is to isolate the effect of terrorism on the GDP of Basque Country. In other words, to evaluate the effect of terrorism, SC-like methods aim to estimate the unobservable counterfactual GDP growth in the absence of terrorism for Basque Country using observations from various other Spanish regions, which are assumed to be unaffected by the terrorist activity.

Since we do not have access to the counterfactual realities of the Basque Country GDP post 1970 in the absence of terrorism, we will use the celebrated estimates of [2] as our baseline; this is our chosen "ground-truth" because these counterfactual trajectories of Basque's GDP in the absence of terrorism have been widely accepted in the econometrics community. The resulting synthetic Basque is displayed in Figure 2a.

Classical SC and OLS under missing data. We randomly obfuscate data, ranging from 5-20%, and in Figure 2b we plot the resulting synthetic Basque GDPs predicted via convex regression on the outcome GDP data, i.e., the original SC method *without* auxiliary covariates – the solid blue and orange lines represent the observed and synthetic Basque (predicted by [2]), respectively, while the dashed lines represent the synthetic Basques under varying levels of missing data. As clearly seen from the figure, the original SC method is not robust to sparse observations, which may explain its dependency on auxiliary covariates to learn its model.

Additionally, we construct a synthetic Basque using OLS, i.e., running linear regression without any pre-processing of the donor observations (on only the outcome GDPs). As seen in Figure 2c, OLS clearly overfits to the idiosyncratic noise of the pre-intervention data and fails to produce sensible post-intervention estimates. In fact, the synthetic Basque GDP as predicted by OLS suggests terrorism actually had a long-term benefit for the Basque economy! This example motivates the importance of appropriately regularizing and de-noising the donor data, as PCR does, prior to learning a synthetic control.

Importance of covariate pre-processing via PCA. The first step of PCR (i.e., PCA) is even more starkly empirically motivated by inspecting the singular value spectrum and cumulative energy of the Basque dataset, which are shown in Figures 3a and 3b, respectively. The data exhibits low-dimensional structure with over 99% of the spectral energy captured in just the *top* singular value, which fits the setting under which our theoretical results imply low pre- and post-intervention prediction errors. The resulting synthetic Basque as per PCR (or, equivalently, the RSC method) is shown in Figure 4a, which pleasingly closely matches that of [2]. Similarly, in Figure 4b, we display various synthetic Basque GDPs after randomly obfuscating the donor observations. Across the varying levels of missing data from 5-20%, the synthetic Basque GDPs continue to resemble the baseline estimates of [2] such that the same conclusion on the negative economic effects of terrorism can be drawn.

Importantly, we underscore that all of the results computed via the PCR method shown in Figures 4a and 4b only use the outcome data (only the per-capita GDP values), i.e., the PCR estimator does *not* utilize the auxiliary covariate information that was required to achieve the results in [2]. Hence, PCR exhibits desirable robustness properties with respect to missing and noisy data, and with less stringent data requirements to achieve similar counterfactual estimates.

California Proposition 99. Another popular case study in the SC literature investigates the impact of California's Proposition 99, an anti-tobacco legislation, on the per-capita cigarette consumption in the state (see [1]). Similar to the Basque example, we will use the widely accepted counterfactual estimate of [1] as our baseline, which is shown in Figure 6a. Here, the authors of [1] considered California as the target state, the collection of states in the U.S. that did not adopt some variant of a tobacco control program as the donor pool, and Proposition 99 (enacted in 1988) as the intervention.

We again plot the resulting Californias learned via convex regression *without* auxiliary covariates and under varying levels of missing data (5-20%) in Figure 6b; similar to the Basque case study, this highlights the poor performance of the original SC method in the presence of missing data. Further, we plot the singular value spectrum and energy of the California Proposition 99 dataset, seen in Figures 5a and 5b, and observe that over 99% of the cumulative spectral energy is again captured by the top singular value, which fits the setting under which our theoretical results apply and motivates the application of PCR.

Empirically, we observe that the resulting synthetic California predicted via PCR, also displayed in Figure 6a, closely matches the baseline, again without using any of the auxiliary covariates considered in the work of [1]. This is indeed expected from the theoretical analysis given the extremely low-dimensional structure of the data. Much like the previous Basque example, across the varying levels of missing data from 5-20%, the synthetic California per-capita cigarette consumption trajectories continue to mirror the baseline estimates of [1]—even in the presence of missing data, the counterfactual estimates produced by PCR suggest that Proposition 99 successfully cut smoking in California.

5 Conclusion

Summary of contributions. As the main contribution of this work, we address a long-standing problem of showing PCR (as is) is surprisingly robust to a wide array of problems that plague large-scale modern datasets, including high-dimensional and noisy, sparse, and mixed valued covariates. We provide meaningful non-asymptotic bounds for both the training and testing (transductive semi-supervised setting) errors for these settings, even when the covariate matrix is only *approximately* low-rank and the linear model is *misspecified*. From a practical standpoint, our testing error bound further provides guidance as to how to choose the PCR hyper-parameter k in a data-driven manner. To achieving our formal results, we establish a simple, but powerful equivalence between PCR and linear regression with covariate pre-processing via HSVT; in the process, we provide a novel error analysis of HSVT with respect to the $\ell_{2,\infty}$-norm. We then formally connect our theoretical results with three important applications to highlight the broad meaning of “noisy” covariates; namely, SC (measurement noise), differentially private regression (noise added by design), and mixed covariate regression ("structural" noise). Of particular note, given the equivalence between PCR and the RSC estimator, it immediately leads to a finite-sample bound for the post-intervention error of RSC under a generalized factor model, which is currently absent from the literature. We note that finite-sample analyses are absent for most SC estimators.

How to "robustify" an estimator. In essence, this work shows that the PCA component of PCR is an effective pre-processing tool in finding a linear low-dimensional embedding of the covariates, which carries the added benefits of implicit de-noising and ℓ_0-regularization. We postulate that when the covariate data is "unstructured" (e.g., speech or video), finding meaningful nonlinear low-dimensional embeddings of the data can also achieve similar implicit benefits, e.g., via a variational auto-encoder or a general adversarial network. We hope this work motivates a general statistical principle that to "robustify" a statistical estimator—first find a low-dimensional embedding of the data before fitting a prediction model.

References

[1] 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.

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

[3] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3-4):195–200, 2011.

[4] A. Agarwal, D. Shah, and D. Shen. On principal component regression in a high-dimensional error-in-variables setting, 2020.

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

[6] D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager. Synthetic difference in differences. *arXiv e-prints arXiv:1812.09970.*, 2018.

[7] S. Athey, M. Bayati, N. Doudchenko, and G. Imbens. Matrix completion methods for causal panel data models. 2017.

[8] S. Athey and G. Imbens. The state of applied econometrics - causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2016.

[9] 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.

[10] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, Mar. 2003.

[11] 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*, 2017.

[12] 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, 2017.

[13] E. Ben-Michael, A. Feller, and J. Rothstein. The augmented synthetic control method, 2018.

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

[15] E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[16] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[17] 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.

[18] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

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

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

[21] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[22] N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Paper No. 22791*, 2016.

[23] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[24] C. Hsiao, H. Steve Ching, and S. Ki Wan. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740, 2012.

[25] C. Hsiao, S.-K. Wan, and Y. Xie. Panel data approach vs synthetic control method. *Economics Letters*, 164:121–123, 2018.

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

[27] 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.

[28] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[29] K. T. Li. Inference for factor model based average treatment effects. *Available at SSRN 3112775*, 2018.

[30] K. T. Li and D. R. Bell. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197(1):65 – 75, 2017.

[31] 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.

[32] R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

[33] H. R. Moon and M. Weidner. Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579, 2015.

[34] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *J. Mach. Learn. Res.*, 11:2241–2259, Aug. 2010.

[35] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[36] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 04 2011.

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

[38] 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.

[39] D. Shah and D. Song. Learning mixture model with missing values and its application to rankings. *arXiv preprint arXiv:1812.11917*, 2018.

[40] 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.

[41] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Publishing Company, Incorporated, 2008.

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

[43] M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

[44] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

[46] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[47] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[48] J. Xu. Rates of convergence of spectral methods for graphon estimation. *arXiv preprint arXiv:1709.03183*, 2017.

[49] Y. Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Econometrics: Multiple Equation Models eJournal*, 2016.

6 Figures

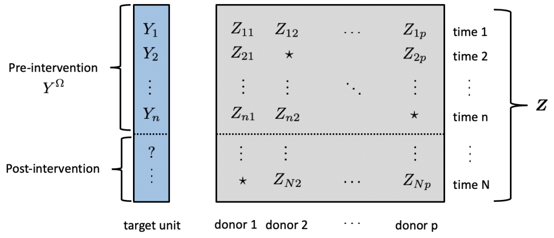
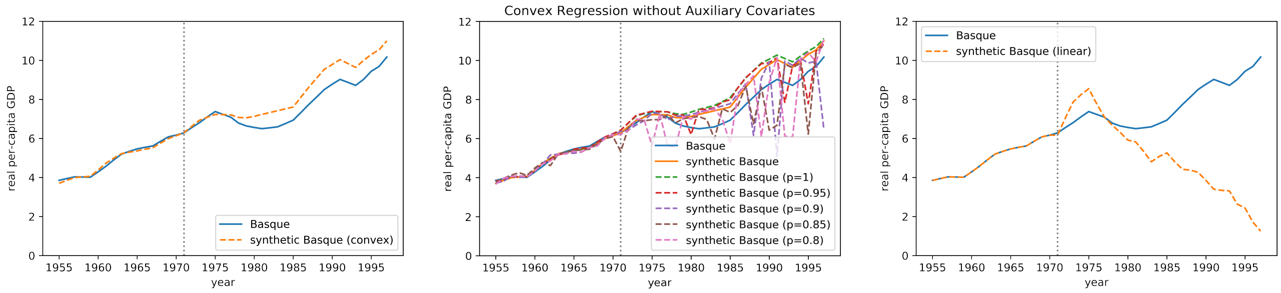


Figure 1: Caricature of observed data (Y^{pre},$\boldsymbol{Z}$) in SC framework (with ⋆ denoting unobserved and/or missing data in the donor matrix). "?" represents the counterfactual observations for the target unit in the absence of intervention, which is what we wish to estimate.

Figure 2:

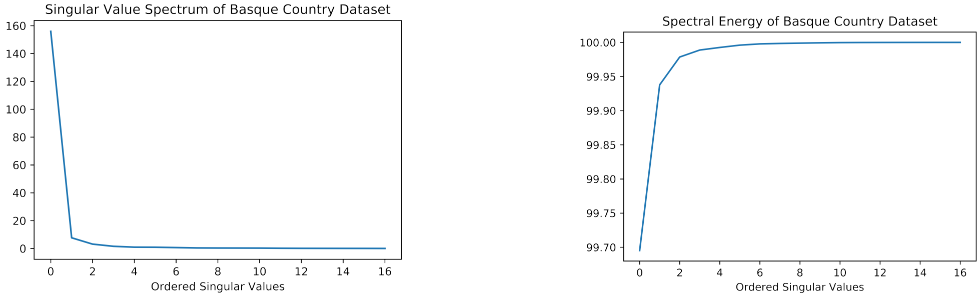


(a) Synthetic Basque as predicted by [2].

(b) Synthetic Basque as predicted by [2] under varying levels of missing data.

(c) Synthetic Basque as predicted by Linear Regression.

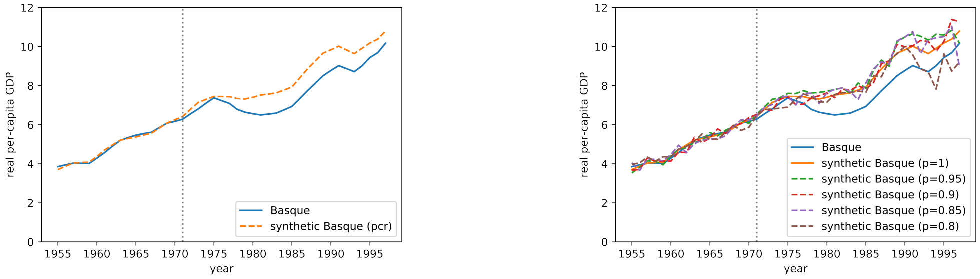
Figure 3:



(a) Singular value spectrum of Basque Country dataset.

(b) Spectral energy of Basque Country dataset.

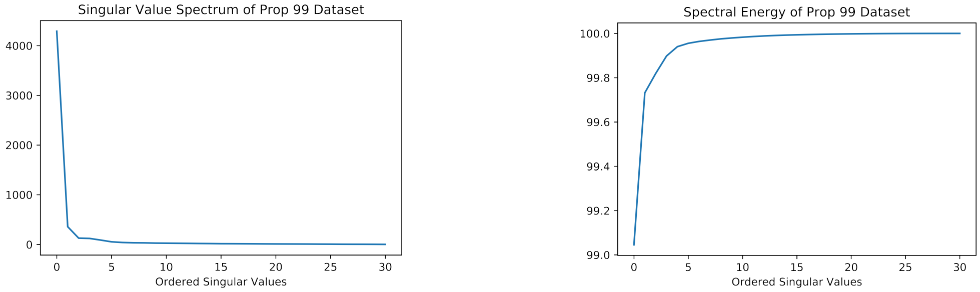
Figure 4:



(a) Synthetic Basque as predicted by PCR.

(b) Synthetic Basque as predicted by PCR under varying levels of missing data.

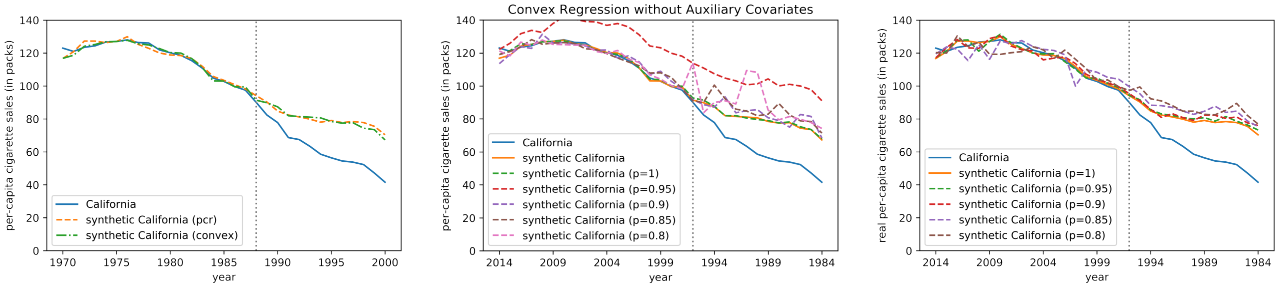
Figure 5:



(a) Singular value spectrum of California Prop 99 dataset.

(b) Spectral energy of California Prop 99 dataset.

Figure 6:



(a) Synthetic California as predicted by PCR and [1].

(b) Synthetic California as predicted by [1] under varying levels of missing data.

(c) Synthetic California as predicted by PCR under varying levels of missing data.

Online Supplement:
On Robustness of Principal Component Regression

Anish Agarwal, Devavrat Shah, Dennis Shen, Dogyoon Song

MIT

A Related Works

We focus on the related literature pertaining to error-in-variable regression and PCR, but also include a brief discussion on the literature for matrix estimation/completion.

Error-in-variables regression. There exists a rich body of work regarding high-dimensional error-in-variable regression (see [31], [20], [37], [19], [27]). Three common threads of these works include: (1) making a sparsity assumption on β^*; (2) establishing error bounds with convergence rates for estimating β^* under different norms, i.e., $\|\widehat{\beta}-\beta^*\|_q$ where $\|\cdot\|_q$ denotes the ℓ_q-norm; (3) assuming the covariate matrix satisfying "incoherence"-like condition such as the Restricted Eigenvalue Condition, cf. [31]. In all of these works, the goal is to recover the underlying model, β^*. In contrast, as discussed, the goal of PCR is to primarily provide good prediction. Some notable works closest to our setup include [31], [20], [38], which are described in some detail next.

In [31], a non-convex ℓ_1-penalization algorithm is proposed based on the plug-in principle to handle covariate measurement errors. This approach requires explicit knowledge of the unobserved noise covariance matrix $\Sigma_{\boldsymbol{H}} = \mathbb{E}\boldsymbol{H}^T\boldsymbol{H}$ and the estimator designed *changes* based on their assumption of $\Sigma_{\boldsymbol{H}}$. They also require explicit knowledge of a bound on the $\|\cdot\|_2$-norm of β^*, the object they aim to estimate. In contrast, PCR does not require any such knowledge about the distribution of the noise matrix $\boldsymbol{H}$ (i.e., the algorithm does not explicitly use this information to make predictions).

The work of [20] builds upon [31] by proposing a convex formulation of Lasso. Although the algorithm introduced does not require knowledge of $\|\beta^*\|_2$, similar assumptions on $\boldsymbol{Z}$ and $\boldsymbol{H}$ (e.g., sub-gaussianity and access to $\Sigma_{\boldsymbol{H}}$) are made. This renders their algorithm to not be noise-model agnostic. In fact, many works (e.g., [37], [38], [11]) require either $\Sigma_{\boldsymbol{H}}$ to be known or the structure of $\boldsymbol{H}$ is such that it admits a data-driven estimator for its covariance matrix. This is so because these algorithms rely on correcting the bias for the matrix $\boldsymbol{Z}^T\boldsymbol{Z}$, which PCR does not need to compute.

It is worth noting that all these works in error-in-variables regression focus only on parameter estimation (i.e., learning β^*) and not explicitly de-noising the noisy covariates. Thus, even with the knowledge of β^*, it is not clear how these methods can be used to produce predictions of the response variables associated with unseen, noisy covariates.

Principal Component Regression. A notable work is that of [9], which suggests a variation of PCR to infer the direction of the principal components. However, it stops short of providing meaningful finite sample analysis beyond what is naturally implied by that of standard Linear Regression. The regularization property of PCR is also well known, at least empirically, due to its ability to reduce the variance. As a contribution, we provide rigorous finite sample guarantees of PCR: (i) under noisy, missing covariates; (ii) when the linear model is misspecified; (iii) when the low-rank model for covariate matrix is misspecified.

As a further contribution, we argue that PCR's regression model has sparse support (established using the equivalence between PCR and Linear Regression with covariate pre-processing via HSVT); this sparsity allows for improved generalization as the Rademacher complexity of the resulting model class scales with the sparsity parameter (i.e., the rank of the covariate matrix pre-processed with HSVT). Hence, PCR not only addresses the challenge of noisy, missing covariates, but also, in effect, performs implicit regularization.

Matrix estimation. Matrix estimation has spurred tremendous theoretical and empirical research across numerous fields (see [16, 28, 35, 18]), Traditionally, the end goal is to recover the underlying mean matrix from an incomplete, noisy sampling of its entries; the quality of the estimate is often measured through the Frobenius norm. Further, entry-wise independence and sub-gaussian noise is typically assumed. A key property of many matrix estimation methods is they are noise-model agnostic (i.e., the de-noising procedure does not change with the noise assumptions). We advance state-of-art for HSVT, arguably the most ubiquitous matrix estimation method, by (i) analyzing its error with respect to the $\ell_{2,\infty}$-norm and (ii) allowing for a broader class of noise distributions (e.g., sub-exponential). Such generalizations are necessary to enable the various applications detailed in Section 4 and Appendices B and C.

B Differentially Private Regression

Setup and Question. With the increasing use of machine learning for critical operations, analysts must maximize the accuracy of their predictions and simultaneously protect sensitive information (i.e., covariates). An important notion of privacy is that of differential privacy; this requires that the outcome of a database query cannot greatly change due to the presence or absence of any individual data record (see [23] and references therein). More specifically, let δ be a positive real number, $\mathcal{D}$ be a collection of datasets, and $\mathcal{A}:\mathcal{D}\rightarrow \mathrm{im}(\mathcal{A})$ be a randomized algorithm that takes a dataset as input. The algorithm $\mathcal{A}$ is said to provide δ-differential privacy if, for all datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ in $\mathcal{D}$ that differ on a single element, and all subsets $\mathcal{S}\in \mathrm{im}(\mathcal{A})$, the following holds:

$$\mathbb{P}(\mathcal{A}(D_1)\in S)\leq \exp(\delta)\cdot \mathbb{P}(\mathcal{A}(D_2)\in S), \tag{20}$$

where the randomness lies in the algorithm. Thus, (20) guarantees that little can be learned about any particular record within the database.

The canonical mechanism $\mathcal{A}$ to guarantee differential privacy is known as the Laplacian mechanism. In this setting, noise is drawn from a Laplacian distribution and added to query responses. In particular, introducing additive noise $W \sim \text{Laplace}(0,\Delta_f/\delta)$ to any database query guarantees δ-privacy (see [23] and references therein); here, $\Delta_f = \max_{\mathcal{D}_1,\mathcal{D}_2 \in \mathcal{D}} |f(\mathcal{D}_1) - f(\mathcal{D}_2)|$, where the maximum is taken over all pairs of datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ in $\mathcal{D}$ differing in at most one element, and $f: \mathcal{D} \to \mathbb{R}^d$ is a vector-valued function denoting the true, latent query response. We now describe how PCR can be applied in the context of a differentially private framework.

How it fits our framework. Let $\boldsymbol{A}$ denote the true, fixed database of N sensitive individual records and p covariates. We consider the setting where an analyst is allowed to ask two types of queries of the data: (1) $f_{\boldsymbol{A}}$ - querying for individual data records, i.e., $\boldsymbol{A}_{i,\cdot}$ for $i \in [N]$; (2) f_Y - querying for a linear combination of an individual's covariates, i.e. $\boldsymbol{A}_{i,\cdot}\beta^*$. A typical example would be where $\boldsymbol{A}_{i,\cdot}$ is the genomic information for patient i and $\boldsymbol{A}_{i,\cdot}\beta^*$ denotes patient i's outcome for a clinical study.

In order to provide δ-differential privacy, the Laplacian mechanism will return query responses with additive Laplacian noise. For query type (1), let Z_{ij} for $i \in [N], j \in [p]$ be the returned response; here, $Z_{ij} = A_{ij} + \eta_{ij}$ with probability ρ and $Z_{ij} = \star$ with probability $1-\rho$, where $\eta_{i,\cdot} = [\eta_{ij}]$ for $j \in [p]$ is independent Laplacian noise with the variance parameter proportional to $\Delta_{f_{\boldsymbol{A}}}/\delta$; we note that an auxiliary benefit of our setup is that it allows for a significant fraction of the query response to be masked, in addition to to the Laplacian noise corruption. For query type (2), when an analyst queries for the response variable $\boldsymbol{A}_{i,\cdot}\beta^*$, she observes $Y_i = \boldsymbol{A}_{i,\cdot}\beta^* + \epsilon_i$, where ϵ_i is again independent Laplacian noise with variance parameter proportional to Δ_{f_Y}/δ. We note that the above setup naturally fits our framework since the Laplacian distribution belongs to the family of sub-exponential distributions, i.e., satisfying Property 3.3 with $\alpha = 1$.

Finally, let Y^Ω denote the n noisy observed responses (e.g., corresponding to the outcomes of n patient clinical trials), and let $\boldsymbol{Z}$ denote the noisy observed covariates (e.g., the collection of genomic information of all N patients). Ultimately, the goal in such a setup is to accurately learn in- and out-of-sample global statistics (e.g., having low $\text{MSE}_\Omega(\widehat{Y})$ and $\text{MSE}(\widehat{Y})$, respectively) about the data, while preserving the individual privacy of the users.

Is privacy preserved? Lemma 3.1 demonstrates that the estimated covariate matrix $\boldsymbol{Z}^{\text{HSVT},k}$ via HSVT achieves small average $\|\cdot\|_{2,\infty}$-norm error (column-squared error); hence, for instance, HSVT can accurately learn the *average* age of all patients. However, this does not translate to accurately estimating the age of any particular patient – this would correspond to a small $\|\cdot\|_{\infty}$-norm error. Similarly, Corollary B.1 (stated below), establishes that PCR can estimate the vector $\boldsymbol{A}\beta^*$ well on average, but not any particular element of this vector. We leave it as an open question as to whether or not de-noising the covariate matrix through HSVT can give a $\|\cdot\|_{\infty}$-norm bound.

Results. We now state the following corollary, an instantiation of Corollary 3.1, which demonstrates the efficacy of PCR (with respect to prediction) in the context of differential privacy. We note a similar bound could easily be produced for any of the results in Section 3 – see (4), (21), (6), (7), (8), (12), (14) – by appropriately substituting γ, K_α with $\frac{\Delta_{f_{\boldsymbol{A}}}}{\delta}$ and σ with $\frac{\Delta_{f_Y}}{\delta}$.

Corollary B.1. *Let the conditions of Corollary 3.1. Let η_{ij} be sampled independently from $\sim \text{Laplace}(0, \Delta_{f_{\boldsymbol{A}}}/\delta)$ for $i \in [N], j \in [p]$. Let ϵ_i be sampled independently from $\sim \text{Laplace}(0, \Delta_{f_Y}/\delta)$. Let $n = \Theta(N)$. Then, PCR preserves δ-differential privacy of $\boldsymbol{A}$ and $\boldsymbol{A}\beta^*$ with*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{C'\|\beta^*\|_1^2}{\rho^4} \frac{r\log^5(np)}{n \wedge p} + \frac{20\|\phi\|_2^2}{n}, \tag{21}$$

where $C' = C(1 + (\Delta_{f_Y}/\delta)^2)(1 + (\Delta_{f_{\boldsymbol{A}}}/\delta)^8)$ and $C > 0$ is an absolute constant.

Proof. Proof is immediate from Corollary 3.1 by substituting γ, K_α with $\frac{\Delta_{f_{\boldsymbol{A}}}}{\delta}$ and σ with $\frac{\Delta_{f_Y}}{\delta}$. □

Interpretation. From Corollary B.1, we observe that PCR learns a predictive linear model in a differentially private framework, where the covariates are purposefully contaminated with Laplacian noise to maintain δ-differential privacy.

C Regression with Mixed Valued Covariates

Setup and Question. Regression models with mixed discrete and continuous covariates are ubiquitous in practice. With respect to discrete covariates, a standard generative model assumes the covariates are generated from a categorical distribution (i.e., a multinomial distribution). Formally, a categorical distribution for a random variable X is such that X has support in $[G]$ and the probability mass function (pmf) is given by $\mathbb{P}(X=g)=\rho_g$ for $g\in[G]$ with $\sum_{g=1}^{G}\rho_g=1$.

For simplicity, we focus on the case where the regression is being done with a collection of Bernoulli random variables (i.e., each X has support in $\{0,1\}$). The extension to general categorical random variables is straightforward and discussed below.

A standard model in regression with Bernoulli random variables assumes that the response variable is a linear function of the latent parameters of the observed discrete outcomes. Formally, $\boldsymbol{A}_{i,\cdot} = [\rho_1^{(i)}, \rho_2^{(i)}, \dots, \rho_p^{(i)}] \in \mathbb{R}^{1\times p}$, where $\rho_j^{(i)}$ for $j \in [p]$ is the latent Bernoulli parameter for the j-th feature and i-th measurement. Further, the mean of the response variable satisfies $\mathbb{E}[Y_i]=\sum_{j=1}^{p}\rho_j^{(i)}\beta_j$. However, for each feature, we only get binary observations, i.e., $X_{ij}\in\{0,1\}$.

As an example, consider $\mathbb{E}[Y_i]$ to be the expected health outcome of patient i. Let there be a total of p possible observable binary symptoms (e.g., cold, fever, headache, etc.). Then $\boldsymbol{A}_{i,\cdot}$ denotes the vector of (unobserved) probabilities that patient i has some collection of symptoms (e.g., $A_{i1} = \mathbb{P}(\text{patient } i \text{ has a cold}), A_{i2} = \mathbb{P}(\text{patient } i \text{ has a fever}), \dots$). However, for each patient, we only observe the "noisy" binary outcome of these symptoms (i.e., $X_{i1} = \mathbb{1}(\text{patient } i \text{ has a cold}), X_{i2} = \mathbb{1}(\text{patient } i \text{ has a fever})$). Ideally, we get to observe the underlying probabilities of the symptoms as that is what we assume the response is linearly related to. The objective in such a setting is to accurately recover $\boldsymbol{A}\beta^*$ given Y^Ω and $\boldsymbol{X}$.

Current practice for mixed valued features. A common practice for regression with categorical variables is to build a separate regression model for every possible combination of the categorical outcomes (i.e., to build a separate regression model conditioned on each outcome). In the healthcare example above, this would amount to building 2^p separate regression models corresponding to each combination of the observed p binary symptoms. This is clearly not ideal for the following two major reasons: (i) the sample complexity is exponential in p; (ii) we do not have access to the underlying probabilities $\boldsymbol{A}_{i,\cdot}$ (recall $\boldsymbol{X}_{i,\cdot} \in \{0,1\}^p$), which is what we actually want to regress Y^Ω against.

How it fits our framework. Recall from Property 3.3 that the key structure we require of the covariate noise η_{ij} is that $\mathbb{E}[\eta_{ij}] = 0$. Now even though $X_{ij} \in \{0,1\}$, it still holds that $\mathbb{E}[X_{ij}] = \rho_j^{(i)} = A_{ij}$, which immediately implies $\mathbb{E}[\eta_{ij}] = \mathbb{E}[X_{ij} - A_{ij}] = 0$. Further, η_{ij} is sub-Gaussian ($\alpha = 2$) since $|\eta_{ij}| \leq 1$. Thus, the key conditions on the noise are satisfied for PCR to effectively (in the $\|\cdot\|_{2,\infty}$-norm) de-noise $\boldsymbol{X}$ to recover the underlying probability matrix $\boldsymbol{A}$; this, in turn, allows PCR to produce accurate estimates $\widehat{\boldsymbol{A}}\widehat{\beta}$ through regression, as seen by Theorem 3.2.

Pleasingly, the required sample complexity grows with the rank of $\boldsymbol{A}$ (the inherent model complexity of the underlying probabilities), rather than exponentially in p. Further, the de-noising step allows us to regress against the estimated latent probabilities rather than their "noisy", binary outcomes.

Extension from Bernoulli to general categorical random variables. Recall from above that a categorical random variable has support in $[G]$ for $G \in \mathbb{N}$. In this case, one can translate a categorical random variable to a a collection of binary random variables using the standard one-hot encoding method. It is worth highlighting that by using one-hot encoding, clearly η_{ij_1} will not be independent of η_{ij_2} for any (j_1, j_2) pair, which encodes the same categorical variable. However, from Property 3.3, we only require independence of the noise across rows, not within them. Thus this lack of independence is not an issue. Further, the generalization to multiple categorical variables, in addition to continuous covariates, is achieved by simply appending these features to each row and collectively de-noising the entire matrix before the regression step.

D Useful Theorems Known from Literature

D.1 Bounding ψ_α-norm

Lemma D.1. Sum of independent sub-gaussians random variables.

Let $X_1,\ldots,X_n$ be independent, mean zero, sub-gaussian random variables. Then $\sum_{i=1}^n X_i$ is also a sub-gaussian random variable, and

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

where C is an absolute constant.

Lemma D.2. Product of sub-gaussians is sub-exponential.

Let X and Y be sub-gaussian random variables. Then XY is sub-exponential. Moreover,

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

D.2 Concentration Inequalities for Random Variables

Lemma D.3. Bernstein's inequality.

Let $X_1,X_2,\ldots,X_N$ be independent, mean zero, sub-exponential random variables. Let $S=\sum_{i=1}^n X_i$. Then for every $t>0$, we have

$$\mathbb{P}\{|S|\geq t\} \leq 2\exp\left(-c\min\left[\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\Psi_1}^2},\frac{t}{\max_i \|X_i\|_{\Psi_1}}\right]\right)$$

Lemma D.4. McDiarmid inequality.

Let $x_1,\ldots,x_n$ be independent random variables taking on values in a set A, and let $c_1,\ldots,c_n$ be positive real *constants. If* $\phi : A^n \to \mathbb{R}$ *satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} |\phi(x_1,\ldots,x_i,\ldots,x_n) - \phi(x_1,\ldots,x_i',\ldots,x_n)| \le c_i,$$

for $1 \le i \le n$*, then*

$$\mathbb{P}\Big\{|\phi(x_1,\ldots,x_n) - \mathbb{E}\phi(x_1,\ldots,x_n)| \ge \epsilon\Big\} \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

D.2.1 Upper Bound on the Maximum Absolute Value in Expectation

Lemma D.5. Maximum of sequence of random variables.

Let $X_1, X_2, \ldots, X_n$ *be a sequence of random variables, which are not necessarily independent, and satisfy* $\mathbb{E}[X_i^{2p}]^{\frac{1}{2p}} \le K p^{\frac{\beta}{2}}$ *for some* $K, \beta > 0$ *and all* i*. Then, for every* $n \ge 2$*,*

$$\mathbb{E}\max_{i \le n} |X_i| \le CK \log^{\frac{\beta}{2}}(n).$$

Remark D.1. *Lemma D.5 implies that if* $X_1, \ldots, X_n$ *are* ψ_α *random variables with* $\|X_i\|_{\psi_\alpha} \le K_\alpha$ *for all* $i \in [n]$*, then*

$$\mathbb{E}\max_{i \le n} |X_i| \le CK_\alpha \log^{\frac{1}{\alpha}}(n).$$

D.3 Other Useful Lemmas

Lemma D.6. Perturbation of singular values (Weyl's inequality).

Let $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *be two* $m \times n$ *matrices. Let* $k = m \wedge n$*. Let* $\lambda_1, \ldots, \lambda_k$ *be the singular values of* $\boldsymbol{A}$ *in decreasing order and repeated by multiplicities, and let* $\tau_1, \ldots, \tau_k$ *be the singular values of* $\boldsymbol{B}$ *in decreasing order and repeated by multiplicities. Let* $\delta_1, \ldots, \delta_k$ *be the singular values of* $\boldsymbol{A} - \boldsymbol{B}$*, in any order but still repeated by multiplicities. Then,*

$$\max_{1 \le i \le k} |\lambda_i - \tau_i| \le \max_{1 \le i \le k} |\delta_i|.$$

E Definitions

Definition E.1 (ψ_α-random variables/vectors)**.** *For any $\alpha \geq 1$, we define the ψ_α-norm of a random variable X as $\|X\|_{\psi_\alpha} = \inf\{t > 0 : \mathbb{E}\exp(|X|^\alpha / t^\alpha) \leq 2\}$. If $\|X\|_{\psi_\alpha} < \infty$, we call X a ψ_α-random variable. More generally, we say X in $\mathbb{R}^n$ is a ψ_α-random vector if all one-dimensional marginals $\langle X, v\rangle$ are ψ_α-random variables for any fixed vector $v \in \mathbb{R}^n$. We define the ψ_α-norm of the random vector $X \in \mathbb{R}^n$ as $\|X\|_{\psi_\alpha} = \sup_{v \in \mathcal{S}^{n-1}} \|\langle X, v\rangle\|_{\psi_\alpha}$, where $\mathcal{S}^{n-1} := \{v \in \mathbb{R}^n : \|v\|_2 = 1\}$, $\langle \cdot, \cdot \rangle$ usual inner product. Note that $\alpha = 2$ and $\alpha = 1$ represent the class of sub-gaussian and sub-exponential random variables/vectors, respectively.*

Definition E.2 (Restricted Eigenvalue (RE) condition)**.** *For some $\alpha \geq 1$, and non-empty subset $S \in [p]$, let*

$$\mathcal{C}_\alpha(S) = \{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\},$$

where $S^c = [p] \setminus S$ and $\Delta_S = \{\Delta_j : j \in S\}$.

We say that $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ satisfies the RE(α,κ) condition w.r.t. S if

$$\frac{1}{n}\|\boldsymbol{X}\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2, \quad \forall \Delta \in \mathcal{C}_\alpha(S)$$

where $\kappa > 0$.

F Proof of Proposition 2.1: Equivalence between PCR and HSVT-OLS

Proof of Proposition 2.1. Using the orthonormality of $\boldsymbol{U}, \boldsymbol{V}$, we obtain

$$\begin{aligned}
\widehat{Y}^{\mathrm{PCR},k} &= \widetilde{\boldsymbol{Z}} \cdot \boldsymbol{V}_k \cdot \beta^{\mathrm{PCR},k} \\
&= \widetilde{\boldsymbol{Z}} \cdot \boldsymbol{V}_k \cdot \left(\boldsymbol{Z}^{\mathrm{PCR},k,\Omega}\right)^\dagger Y^\Omega \\
&= \boldsymbol{U} \cdot \boldsymbol{S} \cdot \boldsymbol{V}^T \cdot \boldsymbol{V}_k \cdot \left((\widetilde{\boldsymbol{Z}} \cdot \boldsymbol{V}_k)^\Omega\right)^\dagger \cdot Y^\Omega
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \left((\boldsymbol{U}_k \cdot \boldsymbol{S}_k)^{\Omega} \right)^{\dagger} \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \left(\boldsymbol{U}_k^{\Omega} \cdot \boldsymbol{S}_k \right)^{\dagger} \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \boldsymbol{S}_k^{-1} (\boldsymbol{U}_k^{\Omega})^T \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot (\boldsymbol{U}_k^{\Omega})^T \cdot Y^{\Omega}. \qquad (22)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\widehat{Y}^{\mathrm{HSVT},k} &= \boldsymbol{Z}^{\mathrm{HSVT},k} \cdot \beta^{\mathrm{HSVT},k} \; = \; \boldsymbol{Z}^{\mathrm{HSVT},k} \cdot \left(\boldsymbol{Z}^{\mathrm{HSVT},k,\Omega} \right)^{\dagger} \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \boldsymbol{V}_k^T \cdot \left((\boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \boldsymbol{V}_k^T)^{\Omega} \right)^{\dagger} \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \boldsymbol{V}_k^T \cdot \left(\boldsymbol{U}_k^{\Omega} \cdot \boldsymbol{S}_k \cdot \boldsymbol{V}_k^T \right)^{\dagger} \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot \boldsymbol{S}_k \cdot \boldsymbol{V}_k^T \cdot \boldsymbol{V}_k \cdot \boldsymbol{S}_k^{-1} \cdot (\boldsymbol{U}_k^{\Omega})^{\dagger} \cdot Y^{\Omega} \\
&= \boldsymbol{U}_k \cdot (\boldsymbol{U}_k^{\Omega})^T \cdot Y^{\Omega}. \qquad (23)
\end{aligned}
$$

From (22) and (23), we obtain $\widehat{Y}^{\mathrm{PCR},k} = \widehat{Y}^{\mathrm{HSVT},k}$ for any $k \leq N$. □

G Proof of Theorem 3.1

G.1 Background

Recall that the (a,b)-mixed norm of a matrix $\boldsymbol{B} \in \mathbb{R}^{N \times p}$ is defined as

$$
\|\boldsymbol{B}\|_{a,b} = \left(\sum_{j=1}^{p} \|\boldsymbol{B}_{\cdot,j}\|_a^b \right)^{1/b} = \left(\sum_{j=1}^{p} \left(\sum_{i=1}^{N} \boldsymbol{B}_{ij}^a \right)^{b/a} \right)^{1/b}.
$$

We are interested in the $(2,\infty)$-mixed norm, which corresponds to the maximum ℓ_2 column norm:

$$
\|\boldsymbol{B}\|_{2,\infty} = \max_{j \in [p]} \|\boldsymbol{B}_{\cdot,j}\|_2 = \max_{j \in [p]} \left(\sum_{i=1}^{N} \boldsymbol{B}_{ij}^2 \right)^{1/2}.
$$

Lemma G.1. *Let $\boldsymbol{B}$ be a real-valued $n \times p$ matrix and x a real-valued p dimensional vector. Let $q_1, q_2 \in [1,\infty]$ with $1/q_1 + 1/q_2 = 1$. Then,*

$$\|\boldsymbol{B}x\|_2 \leq \|x\|_{q_1} \|\boldsymbol{B}\|_{2,q_2}.$$

Proof. Using Hölder's Inequality, we have

$$\|\boldsymbol{B}x\|_2^2 = \sum_{i=1}^{n} \langle \boldsymbol{B}_{i,\cdot}, x\rangle^2 \leq \|x\|_{q_1}^2 \sum_{i=1}^{n} \|\boldsymbol{B}_{i,\cdot}\|_{q_2}^2 = \|x\|_{q_1}^2 \cdot \|\boldsymbol{B}\|_{2,q_2}^2.$$

□

G.2 Proof of Theorem 3.1

Proof. For simplicity of notation, let us define $\widehat{\boldsymbol{A}} = \boldsymbol{Z}^{\text{HSVT},k}$, $\widehat{\boldsymbol{A}}^{\Omega} = \boldsymbol{Z}^{\text{HSVT},k,\Omega}$. Due to the equivalence between PCR and performing linear regression using $\widehat{\boldsymbol{A}}^{\Omega}$ via Proposition 2.1, for the remainder of the proof we shall focus on linear regression using $\widehat{\boldsymbol{A}}^{\Omega}$.

As per notation in Section 2.1, let $\beta^{\text{HSVT},k}$ be the solution of linear regression using $\widehat{\boldsymbol{A}}^{\Omega}$ and predicted response variables $\widehat{Y}^{\text{HSVT},k} = \boldsymbol{Z}^{\text{HSVT},k}\beta^{\text{HSVT},k}$; for simplicity, we will denote $\widehat{\beta} = \beta^{\text{HSVT},k}$ and $\widehat{Y} = \widehat{Y}^{\text{HSVT},k} = \widehat{\boldsymbol{A}}\widehat{\beta}$. Recall, per our model specification in (1) , $Y^{\Omega} = \boldsymbol{A}^{\Omega}\beta^* + \phi + \epsilon$. Now observe

$$\|\widehat{\boldsymbol{A}}^{\Omega}\widehat{\beta} - Y^{\Omega}\|_2^2 = \|\widehat{\boldsymbol{A}}^{\Omega}\widehat{\beta} - \boldsymbol{A}^{\Omega}\beta^* + \phi\|_2^2 + \|\epsilon\|_2^2 - 2\epsilon^T(\widehat{\boldsymbol{A}}^{\Omega}\widehat{\beta} - \boldsymbol{A}^{\Omega}\beta^*) - 2\epsilon^T\phi. \tag{24}$$

On the other hand, the optimality of $\widehat{\beta}$ (recall that $\widehat{\beta} \in \operatorname{argmin}\|\widehat{\boldsymbol{A}}^{\Omega}\widehat{\beta} - Y^{\Omega}\|_2^2$) yields

$$\begin{aligned}\|\widehat{\boldsymbol{A}}^{\Omega}\widehat{\beta} - Y^{\Omega}\|_2^2 &\leq \|\widehat{\boldsymbol{A}}^{\Omega}\beta^* - Y^{\Omega}\|_2^2 \\ &= \|(\widehat{\boldsymbol{A}}^{\Omega} - \boldsymbol{A}^{\Omega})\beta^* + \phi\|_2^2 + \|\epsilon\|_2^2 - 2\epsilon^T(\widehat{\boldsymbol{A}}^{\Omega} - \boldsymbol{A}^{\Omega})\beta^* - 2\epsilon^T\phi. \end{aligned} \tag{25}$$

Combining (24) and (25) and taking expectations, we have

$$\mathbb{E}\|\widehat{\boldsymbol{A}}^{\Omega}\widehat{\beta} - \boldsymbol{A}^{\Omega}\beta^* + \phi\|_2^2 \leq \mathbb{E}\|(\widehat{\boldsymbol{A}}^{\Omega} - \boldsymbol{A}^{\Omega})\beta^* + \phi\|_2^2 + 2\mathbb{E}[\epsilon^T\widehat{\boldsymbol{A}}^{\Omega}(\widehat{\beta} - \beta^*)]. \tag{26}$$

Let us bound the final term on the right hand side of (26). Under our independence assumptions (ϵ is independent of $\boldsymbol{H}$), observe that

$$\mathbb{E}[\epsilon^T \widehat{\boldsymbol{A}}^\Omega]\beta^* = \mathbb{E}[\epsilon^T]\mathbb{E}[\widehat{\boldsymbol{A}}^\Omega]\beta^* = 0.$$

Recall that $\widehat{\beta} = (\widehat{\boldsymbol{A}}^\Omega)^\dagger Y = (\widehat{\boldsymbol{A}}^\Omega)^\dagger \boldsymbol{A}^\Omega \beta^* + (\widehat{\boldsymbol{A}}^\Omega)^\dagger \epsilon + (\widehat{\boldsymbol{A}}^\Omega)^\dagger \phi$. Using the cyclic and linearity properties of the trace operator (coupled with similar independence arguments), we further have

$$\begin{aligned}
\mathbb{E}[\epsilon^T \widehat{\boldsymbol{A}}^\Omega \widehat{\beta}] &= \mathbb{E}[\epsilon^T \widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger] \boldsymbol{A}^\Omega \beta^* + \mathbb{E}[\epsilon^T \widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger \epsilon] + \mathbb{E}[\epsilon^T (\widehat{\boldsymbol{A}}^\Omega)^\dagger]\phi \\
&= \mathbb{E}[\epsilon]^T \mathbb{E}[\widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger] \boldsymbol{A}^\Omega \beta^* + \mathbb{E}\Big[\mathrm{tr}\Big(\epsilon^T \widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger \epsilon\Big)\Big] + \mathbb{E}[\epsilon]^T \mathbb{E}[(\widehat{\boldsymbol{A}}^\Omega)^\dagger]\phi \\
&= \mathbb{E}\Big[\mathrm{tr}\Big(\widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger \epsilon\epsilon^T\Big)\Big] = \mathrm{tr}\Big(\mathbb{E}[\widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger] \cdot \mathbb{E}[\epsilon\epsilon^T]\Big) \le \sigma^2 \mathbb{E}\Big[\mathrm{tr}\Big(\widehat{\boldsymbol{A}}^\Omega (\widehat{\boldsymbol{A}}^\Omega)^\dagger\Big)\Big] \\
&= \sigma^2 \mathbb{E}[\mathrm{rank}(\widehat{\boldsymbol{A}}^\Omega)] \ \le \ \sigma^2 k,
\end{aligned} \tag{27}$$

where the inequality follows from Property 3.2 and the fact that rank of $\widehat{\boldsymbol{A}}^\Omega$ is at most that of $\widehat{\boldsymbol{A}} = \boldsymbol{Z}^{\mathrm{HSVT},k}$ and which by definition at most k. Consider

$$\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^* + \phi\|_2^2 = \|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2 + \|\phi\|_2^2 + 2\phi^T(\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*). \tag{28}$$

and

$$\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^* + \phi\|_2^2 = \|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + \|\phi\|_2^2 + 2\phi^T((\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*). \tag{29}$$

From (27), (28) and (29), the (26) becomes

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2 \le 2\sigma^2 k + \mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 \\
+ 2\mathbb{E}|\phi^T(\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*)| + 2\mathbb{E}|\phi^T((\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*)|.
\end{aligned} \tag{30}$$

By Cauchy-Schwartz, we have

$$|\phi^T(\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*)| \le \|\phi\|_2 \|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2, \tag{31}$$

$$|\phi^T((\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*)| \leq \|\phi\|_2 \|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2. \tag{32}$$

Using (31) and (32) in (30), we obtain

$$\mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2 \leq 2\sigma^2 k + \mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 2\|\phi\|_2 \mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2$$

$$+ 2\|\phi\|_2 \mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2.$$

Applying Jensen's Inequality then gives

$$\mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2 \leq 2\sigma^2 k + \mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 2\|\phi\|_2 \sqrt{\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2}$$

$$+ 2\|\phi\|_2 \sqrt{\mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2}. \tag{33}$$

Now, let

$$x = \mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2, \;\; y = 2\sigma^2 k + \mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 2\|\phi\|_2 \sqrt{\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2}.$$

Then, (33) can be viewed as $x \leq y + 2\|\phi\|_2\sqrt{x}$ with both $x, y \geq 0$. Therefore, either $x \leq 4\|\phi\|_2\sqrt{x}$ or $x \leq 2y$, i.e., $x \leq 2y + 16\|\phi\|_2^2$. Replacing the values of x, y as above yields

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2 &\leq 4\sigma^2 k + 2\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 4\|\phi\|_2 \sqrt{\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2} + 16\|\phi\|_2^2 \\
&\leq 4\sigma^2 k + 2\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 4\|\phi\|_2^2 + \mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 16\|\phi\|_2^2 \\
&= 4\sigma^2 k + 3\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 + 20\|\phi\|_2^2, \qquad (34)
\end{aligned}$$

where the second inequality uses the fact that for any $a, b \in \mathbb{R}$, $2ab \leq a^2 + b^2$. We now apply Lemma G.1 with $q_1 = 1$ and $q_2 = \infty$ to obtain

$$\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\beta^*\|_2^2 \leq \|\beta^*\|_1^2 \|\boldsymbol{A}^\Omega - \widehat{\boldsymbol{A}}^\Omega\|_{2,\infty}^2.$$

Dividing by n on both sides of (34) gives the desired result:

$$\frac{1}{n}\mathbb{E}\|\widehat{\boldsymbol{A}}^\Omega \widehat{\beta} - \boldsymbol{A}^\Omega \beta^*\|_2^2 \leq \frac{4\sigma^2 k}{n} + \frac{3\|\beta^*\|_1^2}{n}\mathbb{E}\|(\widehat{\boldsymbol{A}}^\Omega - \boldsymbol{A}^\Omega)\|_{2,\infty}^2 + \frac{20\|\phi\|_2^2}{n}.$$

□

H Towards the Proof of Lemma 3.1: Spectral Norm Upper Bound of Random Matrices with Sub-Exponential Rows

Here we state and derive bound on the spectral norm of random matrix whose rows (or columns) are generated independently per ψ_α-distribution for $\alpha \geq 1$. This will be crucial in establishing the required "de-noising" properties of HSVT

Theorem H.1. *Suppose Properties 3.1, 3.3 for some $\alpha \geq 1$ hold. Then for any $\delta_1 > 0$,*

$$\|\boldsymbol{Z} - \rho \boldsymbol{A}\| \leq \sqrt{N(1+\sigma^2)(1+\gamma^2)} + C(\alpha)\sqrt{1+\delta_1}\sqrt{p}(K_\alpha + 1)\Big(1 + (2+\delta_1)\log(Np)\Big)^{\frac{1}{\alpha}}\sqrt{\log(Np)}$$

with probability at least $1 - \frac{2}{N^{1+\delta_1}p^{\delta_1}}$. Here, $C(\alpha)$ is an absolute constant that depends only on α.

The upper bound stated in Theorem H.1 is not the sharpest possible. But they are sufficient for our purposes. Sharp bounds for $\alpha = 1$ and $\alpha \geq 2$ can be found in [3] and [44] for example.

H.1 Helper Lemmas for the Proof of Theorem H.1

We begin by presenting Proposition H.1, which holds for general random matrices $\boldsymbol{W} \in \mathbb{R}^{N \times p}$. We note that this result depends on two quantities: (1) $\left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|$ and (2) $\|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}$ for all $i \in [N]$. We then instantiate $\boldsymbol{W} := \boldsymbol{Z} - \rho\boldsymbol{A}$ and present Lemmas H.1 and H.5, which bound (1) and (2), respectively, for our choice of $\boldsymbol{W}$.

Proposition H.1. *Let $\boldsymbol{W} \in \mathbb{R}^{N \times p}$ be a random matrix whose rows $\boldsymbol{W}_{i,\cdot}$ ($i \in [N]$) are independent ψ_α-random vectors for some $\alpha \geq 1$. Then for any $\delta_1 > 0$,*

$$\|\boldsymbol{W}\| \leq \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + C(\alpha)\sqrt{(1+\delta_1)p}\max_{i\in[N]}\|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}\Big(1 + (2+\delta_1)\log(Np)\Big)^{\frac{1}{\alpha}}\sqrt{\log(Np)}$$

with probability at least $1 - \frac{2}{N^{1+\delta_1}p^{\delta_1}}$. Here, $C(\alpha) > 0$ is an absolute constant that depends only on α.

Proof. We prove the proposition in four steps.

Step 1: picking the threshold value. Let $e_1,...,e_p \in \mathbb{R}^p$ denote the canonical basis[1] of $\mathbb{R}^p$. Observe that $\|\boldsymbol{W}_{i,\cdot}\|_2^2 = \boldsymbol{W}_{i,\cdot}\boldsymbol{W}_{i,\cdot}^T = \sum_{j=1}^p (\boldsymbol{W}_{i,\cdot}e_j)^{2}$[2] . Therefore, for any $t \geq 0$,

$$\begin{aligned}
\mathbb{P}\left\{\|\boldsymbol{W}_{i,\cdot}\|_2^2 > t\right\} &= \mathbb{P}\left\{\sum_{j=1}^p (\boldsymbol{W}_{i,\cdot}e_j)^2 > t\right\} \\
&\overset{(a)}{\leq} \sum_{j=1}^p \mathbb{P}\left\{(\boldsymbol{W}_{i,\cdot}e_j)^2 > \frac{t}{p}\right\} \\
&\leq \sum_{j=1}^p \mathbb{P}\left\{|\boldsymbol{W}_{i,\cdot}e_j| > \sqrt{\frac{t}{p}}\right\} \\
&\overset{(b)}{\leq} 2p\exp\left(-C(\alpha)\left(\frac{t}{p\|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}^2}\right)^{\frac{\alpha}{2}}\right),
\end{aligned}$$

where (a) uses the union bound and (b) follows from the definition of ψ_α-random vector ($C(\alpha)$ is an absolute constant which depends only on $\alpha \geq 1$). Choosing $t = C^{\frac{2}{\alpha}} C(\alpha)^{-\frac{2}{\alpha}} p \|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}^2 (\log(2p))^{\frac{2}{\alpha}}$ for some $C > 1$ gives

$$\mathbb{P}\left\{\|\boldsymbol{W}_{i,\cdot}\|_2^2 > C^{\frac{2}{\alpha}} C(\alpha)^{-\frac{2}{\alpha}} p \|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}^2 (\log(2p))^{\frac{2}{\alpha}}\right\} \leq \left(\frac{1}{2p}\right)^{C-1}.$$

Applying the union bound, we obtain

$$\mathbb{P}\left\{\max_{i\in[N]}\|\boldsymbol{W}_{i,\cdot}\|_2^2 > C^{\frac{2}{\alpha}} C(\alpha)^{-\frac{2}{\alpha}} p \max_{i\in[N]}\|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}^2 (\log(2p))^{\frac{2}{\alpha}}\right\} \leq N\left(\frac{1}{2p}\right)^{C-1}.$$

For $\delta_1 > 0$, we define $C(\delta_1) \triangleq 1 + (2+\delta_1)\log_{2p}(Np)$ and let $C = C(\delta_1)$. Also, we define

$$t_0(\delta_1) \triangleq C(\delta_1)^{\frac{2}{\alpha}} C(\alpha)^{-\frac{2}{\alpha}} p \max_{i\in[N]}\|\boldsymbol{W}_{i,\cdot}\|_{\psi_\alpha}^2 (\log(2p))^{\frac{2}{\alpha}}.$$

We have

$$\mathbb{P}\left\{\max_{i\in[N]}\|\boldsymbol{W}_{i,\cdot}\|_2^2 > t_0(\delta_1)\right\} \leq N\left(\frac{1}{2p}\right)^{(2+\delta_1)\log_{2p}(Np)} = \frac{1}{N^{1+\delta_1}p^{2+\delta_1}}. \tag{35}$$

[1] Column vector representation

[2] Recall that $\boldsymbol{W}_{i,\cdot}$ is a row vector and hence $\boldsymbol{W}_{i,\cdot}\boldsymbol{W}_{i,\cdot}^T$ is a scalar.

Step 2: decomposing $\boldsymbol{W}$ by truncation. Next, given $\delta_1 > 0$, we decompose the random matrix $\boldsymbol{W}$ as follows:

$$\boldsymbol{W} = \boldsymbol{W}^{\circ}(\delta_1) + \boldsymbol{W}^{\times}(\delta_1)$$

where for each $i \in [N]$,

$$\boldsymbol{W}^{\circ}(\delta_1)_{i,\cdot} = \boldsymbol{W}_{i,\cdot}\mathbb{1}\{\|\boldsymbol{W}_{i,\cdot}\|_2^2 \le t_0(\delta_1)\} \quad \text{and} \quad \boldsymbol{W}^{\times}(\delta_1)_{i,\cdot} = \boldsymbol{W}_{i,\cdot}\mathbb{1}\{\|\boldsymbol{W}_{i,\cdot}\|_2^2 > t_0(\delta_1)\}.$$

Then it follows that

$$\|\boldsymbol{W}\| \le \|\boldsymbol{W}^{\circ}(\delta_1)\| + \left\|\boldsymbol{W}^{\times}(\delta_1)\right\| \le \|\boldsymbol{W}^{\circ}(\delta_1)\| + \left\|\boldsymbol{W}^{\times}(\delta_1)\right\|_F. \tag{36}$$

Step 3: bounding $\|\boldsymbol{W}^{\circ}(\delta_1)\|$ and $\|\boldsymbol{W}^{\times}(\delta_1)\|_F$. We define two events for conditioning:

$$E_1(\delta_1) := \left\{ \|\boldsymbol{W}^{\circ}(\delta_1)\| \le \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + \sqrt{\frac{1+\delta_1}{c} t_0(\delta_1)\log(Np)} \right\}, \tag{37}$$

$$E_2(\delta_1) := \left\{ \left\|\boldsymbol{W}^{\times}(\delta_1)\right\|_F = 0 \right\}. \tag{38}$$

First, given $\delta_1 > 0$, we let $\Sigma^{\circ}(\delta_1) = \mathbb{E}\boldsymbol{W}^{\circ}(\delta_1)^T\boldsymbol{W}^{\circ}(\delta_1)$. By definition of $\boldsymbol{W}^{\circ}(\delta_1)$, we have $\|\boldsymbol{W}_{i,\cdot}\|_2 \le \sqrt{t_0(\delta_1)}$ for all $i \in [N]$. Then it follows that for every $s \ge 0$,

$$\|\boldsymbol{W}^{\circ}(\delta_1)\| \le \|\Sigma^{\circ}(\delta_1)\|^{1/2} + s\sqrt{t_0(\delta_1)}$$

with probability at least $1 - p\exp(-cs^2)$ (see Theorem 5.44 of [44] and Eqs. (5.32) and (5.33) in reference, and replacing the common second moment $\boldsymbol{\Sigma} = \mathbb{E}\boldsymbol{W}_{i,\cdot}^T\boldsymbol{W}_{i,\cdot}$ with the average second moment for all rows, $\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\boldsymbol{W}_{i,\cdot}^T\boldsymbol{W}_{i,\cdot}$, i.e., redefining $\boldsymbol{\Sigma}$). Note that $\|\Sigma^{\circ}(\delta_1)\| = \left\|\mathbb{E}\boldsymbol{W}^{\circ}(\delta_1)^T\boldsymbol{W}^{\circ}(\delta_1)\right\| \le \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|$. Now we define $\tilde{E}_1(s)$ parameterized by $s > 0$ as

$$\tilde{E}_1(s;\delta_1) := \left\{ \|\boldsymbol{W}^{\circ}(\delta_1)\| > \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + s\sqrt{t_0(\delta_1)} \right\}.$$

If we pick $s = \left(\frac{1+\delta_1}{c}\log(Np)\right)^{1/2}$, then $E_1(\delta_1) = \tilde{E}_1(s;\delta_1)$ and

$$\mathbb{P}(E_1(\delta_1)^c) \le p\exp\left(-cs^2\right) = p\exp(-(1+\delta_1)\log(Np)) = \frac{1}{N^{1+\delta_1}p^{\delta_1}}.$$

Next, we observe that $\|\boldsymbol{W}^{\times}(\delta_1)\|_F = 0$ if and only if $\boldsymbol{W}^{\times}(\delta_1) = 0$. If $\boldsymbol{W}^{\times}(\delta_1) \neq 0$, then $\max_{i\in[n]}\|\boldsymbol{W}_{i,\cdot}\|_2^2 > t_0(\delta_1)$. Therefore,

$$\mathbb{P}(E_2^c) \le \frac{1}{N^{1+\delta_1}p^{2+\delta_1}}$$

by the analysis in Step 1; see (35).

Step 4: concluding the proof. For any given $\delta_1 > 0$,

$$\mathbb{P}\left(\|\boldsymbol{W}\| > \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + \sqrt{\frac{1+\delta_1}{c}t_0(\delta_1)\log(Np)} \;\middle|\; E_1(\delta_1)\cap E_2(\delta_1) \right) = 0.$$

by (36), (37), and (38). By the law of total probability and the union bound,

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{W}\| > \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + \sqrt{\frac{1+\delta_1}{c}t_0(\delta_1)\log(Np)} \right) \\
&\quad \le \mathbb{P}\left(\|\boldsymbol{W}\| > \left\|\mathbb{E}\boldsymbol{W}^T\boldsymbol{W}\right\|^{1/2} + \sqrt{\frac{1+\delta_1}{c}t_0(\delta_1)\log(Np)} \;\middle|\; E_1(\delta_1)\cap E_2(\delta_1) \right) \\
&\qquad + \mathbb{P}(E_1(\delta)^c) + \mathbb{P}(E_2(\delta)^c) \\
&\quad \le \frac{1}{N^{1+\delta_1}p^{\delta_1}} + \frac{1}{N^{1+\delta_1}p^{2+\delta_1}} \\
&\quad \le \frac{2}{N^{1+\delta_1}p^{\delta_1}}.
\end{aligned}$$

This completes the proof. □

H.2 Lemmas H.1 and H.5

H.2.1 Lemma H.1

Lemma H.1.

$$\left\|\mathbb{E}(\boldsymbol{Z}-\rho\boldsymbol{A})^T(\boldsymbol{Z}-\rho\boldsymbol{A})\right\| \le \rho(1-\rho)\left(\max_{j\in[p]}\|\boldsymbol{A}_{\cdot,j}\|_2^2 + \|\mathrm{diag}(\mathbb{E}[\boldsymbol{H}^T\boldsymbol{H}])\|\right) + \rho^2\left\|\mathbb{E}\boldsymbol{H}^T\boldsymbol{H}\right\|.$$

Proof. We follow the proof of Lemma A.2 of [39] and state it here for completeness. Throughout, for any matrix $\boldsymbol{Q} \in \mathbb{R}^{N\times p}$, let $Q_\ell \in \mathbb{R}^n$ denote the ℓ-th row of $\boldsymbol{Q}$.

To begin, observe that

$$\mathbb{E}[(\boldsymbol{Z}-\rho\boldsymbol{A})^T(\boldsymbol{Z}-\rho\boldsymbol{A})] = \sum_{\ell=1}^{N}\mathbb{E}[(Z_\ell-\rho A_\ell)\otimes(Z_\ell-\rho A_\ell)].$$

Let $\boldsymbol{X}=\boldsymbol{A}+\boldsymbol{H}$. Importantly, we highlight the following relations: for any $(\ell,i)\in[N]\times[p]$,

$$\mathbb{E}[Z_{\ell i}] = \rho A_{\ell i}$$

$$\mathbb{E}[Z_{\ell i}^2] = \rho\mathbb{E}[X_{\ell i}^2].$$

Now, let us fix a row $\ell \in [N]$ and denote

$$\boldsymbol{W}^{(\ell)} = (Z_\ell-\rho A_\ell)\otimes(Z_\ell-\rho A_\ell).$$

Using the linearity of expectations, the expected value of the (i,j)-th entry of $\boldsymbol{W}^{(\ell)}$ can be written as

$$\mathbb{E}[W_{ij}^{(\ell)}] = \mathbb{E}[Z_{\ell i}Z_{\ell j}] - \rho\mathbb{E}[Z_{\ell i}A_{\ell j}] - \rho\mathbb{E}[Z_{\ell j}A_{\ell i}] + \rho^2\mathbb{E}[A_{\ell i}A_{\ell j}].$$

Suppose $i=j$, then

$$\mathbb{E}[W_{ii}^{(\ell)}] = \rho\mathbb{E}[X_{\ell i}^2] - \rho^2 A_{\ell i}^2 = \rho(1-\rho)\mathbb{E}[X_{\ell i}^2] + \rho^2\mathbb{E}[(X_{\ell i}-A_{\ell i})^2]. \tag{39}$$

On the other hand, if $i \neq j$,

$$\mathbb{E}[W_{ij}^{(\ell)}] = \rho^2\mathbb{E}[(X_{\ell i}-A_{\ell i})(X_{\ell j}-A_{\ell j})]. \tag{40}$$

Therefore, we can express $\boldsymbol{W}^{(\ell)}$ as the sum of two matrices where the diagonal components are generated from (39) and the off-diagonal components are generated from (40). That is,

$$\mathbb{E}[\boldsymbol{W}^{(\ell)}] = \mathbb{E}\Big(\rho(1-\rho)\text{diag}(X_\ell\otimes X_\ell) + \rho^2\text{diag}(H_\ell\otimes H_\ell)\Big) + \mathbb{E}\Big(\rho^2(H_\ell\otimes H_\ell) - \rho^2\text{diag}(H_\ell\otimes H_\ell)\Big)$$

$$= \rho(1-\rho)\mathbb{E}[\mathrm{diag}(X_\ell \otimes X_\ell)] + \rho^2 \mathbb{E}[H_\ell \otimes H_\ell].$$

Taking the sum over all rows $\ell \in [N]$ yields

$$\mathbb{E}[(\boldsymbol{Z} - \rho \boldsymbol{A})^T (\boldsymbol{Z} - \rho \boldsymbol{A})] = \rho(1-\rho)\mathrm{diag}(\mathbb{E}[\boldsymbol{X}^T \boldsymbol{X}]) + \rho^2 \mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}]. \tag{41}$$

To complete the proof, we apply triangle inequality to (41) to obtain

$$\left\| \mathbb{E}[(\boldsymbol{Z} - \rho \boldsymbol{A})^T (\boldsymbol{Z} - \rho \boldsymbol{A})] \right\| \leq \rho(1-\rho) \left\| \mathrm{diag}(\mathbb{E}[\boldsymbol{X}^T \boldsymbol{X}]) \right\| + \rho^2 \left\| \mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}] \right\|.$$

Since $\boldsymbol{H}$ is zero mean, we have

$$\begin{aligned} \left\| \mathrm{diag}(\mathbb{E}[\boldsymbol{X}^T \boldsymbol{X}]) \right\| &= \left\| \mathrm{diag}(\boldsymbol{A}^T \boldsymbol{A}) + \mathrm{diag}(\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}]) \right\| \\ &\leq \left\| \mathrm{diag}(\boldsymbol{A}^T \boldsymbol{A}) \right\| + \left\| \mathrm{diag}(\mathbb{E}[\boldsymbol{H}^T \boldsymbol{H}]) \right\|. \end{aligned}$$

Collecting terms completes the proof. □

H.3 Lemma H.5

Lemma H.2. *Suppose that $X \in \mathbb{R}^n$ and $P \in \{0,1\}^n$ are random vectors. Then for any $\alpha \geq 1$,*

$$\|X \circ P\|_{\psi_\alpha} \leq \|X\|_{\psi_\alpha}.$$

Proof. Given a deterministic binary vector $P_0 \in \{0,1\}^n$, let $I_{P_0} = \{i \in [n] : Q_i = 1\}$. Observe that

$$X \circ P_0 = \sum_{i \in I_{P_0}} e_i e_i^T X.$$

Here, $\circ$ denotes the Hadamard product (entrywise product) of two matrices. By definition of the ψ_α-norm,

$$\|X\|_{\psi_\alpha} = \sup_{u \in \mathbb{S}^{n-1}} \left\| u^T X \right\|_{\psi_\alpha} = \sup_{u \in \mathbb{S}^{n-1}} \inf \left\{ t > 0 : \mathbb{E}_X \left[\exp(|u^T X|^\alpha / t^\alpha) \right] \leq 2 \right\}.$$

Let $u_0 \in \mathbb{S}^{n-1}$ denote the maximum-achieving unit vector (such u_0 exists because $\inf\{\cdots\}$ is continuous with respect to u and $\mathbb{S}^{n-1}$ is compact). Then,

$$\begin{aligned}
\|X \circ P\|_{\psi_\alpha} &= \sup_{u \in \mathbb{S}^{n-1}} \left\|u^T X \circ P\right\|_{\psi_\alpha} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\Big\{t > 0 : \mathbb{E}_{X,P}\Big[\exp\big(|u^T X \circ P|^\alpha / t^\alpha\big)\Big] \le 2\Big\} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\Big\{t > 0 : \mathbb{E}_P\Big[\mathbb{E}_X\Big[\exp\big(|u^T X \circ P|^\alpha / t^\alpha\big) \;\Big|\; P\Big]\Big] \le 2\Big\} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\left\{t > 0 : \mathbb{E}_P\left[\mathbb{E}_X\left[\exp\left(\left|u^T \sum_{i \in I_P} e_i e_i^T X\right|^\alpha / t^\alpha\right) \;\middle|\; P\right]\right] \le 2\right\} \\
&= \sup_{u \in \mathbb{S}^{n-1}} \inf\left\{t > 0 : \mathbb{E}_P\left[\mathbb{E}_X\left[\exp\left(\left|\Big(\sum_{i \in I_P} e_i e_i^T u\Big)^T X\right|^\alpha / t^\alpha\right) \;\middle|\; P\right]\right] \le 2\right\}.
\end{aligned}$$

For any $u \in \mathbb{S}^{n-1}$ and $P_0 \in \{0,1\}^n$, observe that

$$\mathbb{E}_X\left[\exp\left(\left|\Big(\sum_{i \in I_P} e_i e_i^T u\Big)^T X\right|^\alpha / t^\alpha\right) \;\middle|\; P = P_0\right] \le \mathbb{E}_X\Big[\exp\Big(|u_0^T X|^\alpha / t^\alpha\Big)\Big].$$

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

$$\|X \circ P\|_{\psi_\alpha} \le \|X\|_{\psi_\alpha}.$$

□

Lemma H.3. *Let X be a mean-zero, ψ_α-random variable for some $\alpha \ge 1$. Then for $|\lambda| \le \frac{1}{C\|X\|_{\psi_\alpha}}$,*

$$\mathbb{E}\exp(\lambda X) \le \exp\Big(C\lambda^2 \|X\|_{\psi_\alpha}^2\Big).$$

Proof. See [45], Section 2.7. □

Lemma H.4. *Let $X_1, \ldots, X_n$ be independent random variables with mean zero. For $\alpha \ge 1$,*

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_\alpha} \le C\left(\sum_{i=1}^n \|X_i\|_{\psi_\alpha}^2\right)^{1/2}.$$

Proof. Immediate by Lemma H.3. □

Lemma H.5. *Assume Properties 3.1 and 3.3 hold. Then for any $\alpha \geq 1$ with which Property 3.3 holds, we have*

$$\|\boldsymbol{Z}_{i,\cdot} - \rho \boldsymbol{A}_{i,\cdot}\|_{\psi_\alpha} \leq C(K_\alpha + 1) \qquad \text{for all } \; i \in [N],$$

where $C > 0$ is an absolute constant.

Proof. Let $\boldsymbol{P} \in \{0,1\}^{N \times p}$ denote a random matrix whose entries are i.i.d. random variables that take value 1 with probability ρ and 0 otherwise. Note that $\boldsymbol{Z}_{i,\cdot}$ can be written as $\boldsymbol{X}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot}$ where $\star$ is identified with 0. By triangle inequality,

$$\begin{aligned}
\|\boldsymbol{Z}_{i,\cdot} - \rho \boldsymbol{A}_{i,\cdot}\|_{\psi_\alpha} &= \|\boldsymbol{X}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot} - \rho \boldsymbol{A}_{i,\cdot}\|_{\psi_\alpha} \\
&= \|(\boldsymbol{X}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot}) - (\boldsymbol{A}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot}) - \rho \boldsymbol{A}_{i,\cdot} + (\boldsymbol{A}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot})\|_{\psi_\alpha} \\
&\leq \|(\boldsymbol{X}_{i,\cdot} - \boldsymbol{A}_{i,\cdot}) \circ \boldsymbol{P}_{i,\cdot}\|_{\psi_\alpha} + \|(\boldsymbol{A}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot}) - \rho \boldsymbol{A}_{i,\cdot}\|_{\psi_\alpha}.
\end{aligned}$$

By definition of $\boldsymbol{X}$, Property 3.3, and Lemma H.2, we have that

$$\|(\boldsymbol{X}_{i,\cdot} - \boldsymbol{A}_{i,\cdot}) \circ \boldsymbol{P}_{i,\cdot}\|_{\psi_\alpha} \leq \|\boldsymbol{X}_{i,\cdot} - \boldsymbol{A}_{i,\cdot}\|_{\psi_\alpha} = \|\eta_{i,\cdot}\|_{\psi_\alpha} \leq C K_\alpha.$$

Moreover, Property 3.1 and the i.i.d. property of $\boldsymbol{P}_{ij}$ for different j gives

$$\begin{aligned}
\left\|(\boldsymbol{A}_{i,\cdot} \circ \boldsymbol{P}_{i,\cdot}) - \rho \boldsymbol{A}_{i,\cdot}\right\|_{\psi_\alpha} &= \sup_{u \in \mathbb{S}^{p-1}} \left\| \sum_{j=1}^{p} u_j \boldsymbol{A}_{i,j} (\boldsymbol{P}_{i,j} - \rho) \right\|_{\psi_\alpha} \\
&\leq \sup_{u \in \mathbb{S}^{p-1}} \left(\sum_{j=1}^{p} u_j^2 \|\boldsymbol{A}_{i,j} (\boldsymbol{P}_{i,j} - \rho)\|_{\psi_\alpha}^2 \right)^{1/2} \\
&\leq \left(\sup_{u \in \mathbb{S}^{p-1}} \sum_{j} u_j^2 \max_{j \in [p]} |\boldsymbol{A}_{i,j}|^2 \right)^{1/2} \|\boldsymbol{P}_{1,1} - \rho\|_{\psi_\alpha} \\
&\leq \|\boldsymbol{P}_{1,1} - \rho\|_{\psi_\alpha}.
\end{aligned}$$

The first inequality follows from Lemma H.4, the second inequality is immediate, and the last inequality follows from Property 3.1. Lastly, $\|\boldsymbol{P}_{1,1} - \rho\|_{\psi_\alpha} \leq C$ because $\boldsymbol{P}_{1,1} - \rho$ is a bounded random variable in $[-\rho, 1-\rho]$. □

H.4 Proof of Theorem H.1

Proof of Theorem H.1. The proof follows by plugging the results of Lemmas H.1 and H.5 into Proposition H.1 for $\boldsymbol{W} := \boldsymbol{Z} - \rho \boldsymbol{A}$ and applying Properties 3.1 and 3.3. □

I Proof of Lemma 3.1

To bound the error in estimation of HSVT, $\boldsymbol{Z}^{HSVT,k}$ with thresholding at kth singular value, and underlying covariate matrix $\boldsymbol{A}$ with respect to $\|\cdot\|_{2,\infty}$ matrix norm, we shall start by presenting Lemma I.3 which bounds $\|\boldsymbol{Z}^{HSVT,k} - \boldsymbol{A}\|_{2,\infty}$ as a function of few abstract quantities. Next, we bound these quantities with high probability in our setting through help of sequence of results including the spectral norm bound stated in Theorem H.1. We conclude with the proof of Lemma 3.1.

Notation. Consider a matrix $\boldsymbol{B} \in \mathbb{R}^{N \times p}$ such that $\boldsymbol{B} = \sum_{i=1}^{N \wedge p} \sigma_i(\boldsymbol{B}) x_i y_i^T$. With a specific choice of $\lambda \geq 0$, we can define a function $\varphi_\lambda^{\boldsymbol{B}} : \mathbb{R}^N \to \mathbb{R}^N$ as follows: for any vector $w \in \mathbb{R}^N$,

$$\varphi_\lambda^{\boldsymbol{B}}(w) = \sum_{i=1}^{N \wedge p} \mathbb{1}(\sigma_i(\boldsymbol{B}) \geq \lambda) x_i x_i^T w. \tag{42}$$

Note that $\varphi_\lambda^{\boldsymbol{B}}$ is a linear operator and it depends on the tuple $(\boldsymbol{B}, \lambda)$; more precisely, the singular values and the left singular vectors of $\boldsymbol{B}$, as well as the threshold λ. If $\lambda = 0$, then we will adopt the shorthand notation: $\varphi^{\boldsymbol{B}} = \varphi_0^{\boldsymbol{B}}$.

I.1 Lemma I.3

I.1.1 Some Observations on HSVT Operator

Observe that the function $\varphi_\lambda^{\boldsymbol{B}} : \mathbb{R}^N \to \mathbb{R}^N$ defined in (42) is actually the operator acting on the column spaces, which is induced by HSVT.

Lemma I.1. *Let $\boldsymbol{B}\in\mathbb{R}^{N\times p}$ and $\lambda\geq 0$ be given. Then for any $j\in[p]$,*

$$\varphi_{\lambda}^{\boldsymbol{B}}(\boldsymbol{B}_{\cdot,j})=\mathrm{HSVT}_{\lambda}(\boldsymbol{B})_{\cdot,j}.$$

Proof. By (42) and the orthonormality of the left singular vectors,

$$\begin{aligned}\varphi_{\lambda}^{\boldsymbol{B}}(\boldsymbol{B}_{\cdot,j})&=\sum_{i=1}^{N\wedge p}\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda)x_ix_i^T\boldsymbol{B}_{\cdot,j}=\sum_{i=1}^{N\wedge p}\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda)x_ix_i^T\Big(\sum_{i'=1}^{N\wedge p}\sigma_{i'}(\boldsymbol{B})x_{i'}y_{i'}\Big)_{\cdot,j}\\&=\sum_{i,i'=1}^{N\wedge p}\sigma_{i'}(\boldsymbol{B})\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda)x_ix_i^Tx_{i'}(y_{i'})_j=\sum_{i,i'=1}^{N\wedge p}\sigma_{i'}(\boldsymbol{B})\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda)x_i\delta_{ii'}(y_{i'})_j\\&=\sum_{i=1}^{N\wedge p}\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda^*)\sigma_ix_i(y_i)_j\\&=\mathrm{HSVT}_{\lambda}(\boldsymbol{B})_{\cdot,j}.\end{aligned}$$

This completes the proof. □

Remark I.1. *Suppose we have missing data. Then the estimator $\widehat{\boldsymbol{A}}$ has the following representation:*

$$\widehat{\boldsymbol{A}}=\frac{1}{\widehat{\rho}}\mathrm{HSVT}_{\lambda^*}(\boldsymbol{Z})=\frac{1}{\widehat{\rho}}\sum_{i=1}^{N\wedge p}s_i\mathbb{1}(s_i\geq\lambda^*)\cdot u_iv_i^T.$$

By Lemma I.1, we note that

$$\widehat{\boldsymbol{A}}_{\cdot,j}=\frac{1}{\widehat{\rho}}\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{Z}_{\cdot,j}).\tag{43}$$

Lastly, we remark that the column operator induced by HSVT is a contraction.

Lemma I.2. *Let $\boldsymbol{B}\in\mathbb{R}^{N\times p}$ and $\lambda\geq 0$ be given. Then for any $j\in[p]$,*

$$\Big\|\mathrm{HSVT}_{\lambda}(\boldsymbol{B})_{\cdot,j}\Big\|_2\leq\|\boldsymbol{B}_{\cdot,j}\|_2.$$

Proof. By (42) and Lemma I.1, we have

$$\Big\|\mathrm{HSVT}_{\lambda}(\boldsymbol{B})_{\cdot,j}\Big\|_2^2=\big\|\varphi_{\lambda}^{\boldsymbol{B}}(\boldsymbol{B}_{\cdot,j})\big\|_2^2=\Big\|\sum_{i=1}^{N\wedge p}\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda)\cdot x_ix_i^T\cdot\boldsymbol{B}_{\cdot,j}\Big\|_2^2$$

$$\overset{(a)}{=}\sum_{i=1}^{N\wedge p}\left\|\mathbb{1}(\sigma_i(\boldsymbol{B})\geq\lambda)\cdot x_i x_i^T\cdot\boldsymbol{B}_{\cdot,j}\right\|_2^2\leq\sum_{i=1}^{N\wedge p}\left\|x_i x_i^T\cdot\boldsymbol{B}_{\cdot,j}\right\|_2^2$$

$$\overset{(b)}{=}\left\|\sum_{i=1}^{N\wedge p}x_i x_i^T\cdot\boldsymbol{B}_{\cdot,j}\right\|_2^2=\|\boldsymbol{B}_{\cdot,j}\|_2^2.$$

Note that (a) and (b) use the orthonormality of the left singular vectors.

□

Lemma I.3. *Suppose that (1) $\|\boldsymbol{Z}-\rho\boldsymbol{A}\|\leq\Delta$ for some $\Delta\geq0$ and (2) $\frac{1}{\varepsilon}\rho\leq\widehat{\rho}\leq\varepsilon\rho$ for some $\varepsilon\geq1$.*

Let $\widehat{\boldsymbol{A}}=\boldsymbol{Z}^{HSVT,k}$, $\boldsymbol{A}^k=HSVT_{\tau_k}(\boldsymbol{A})$ and $\boldsymbol{E}=\boldsymbol{A}-\boldsymbol{A}^k$. Then for any $j\in[p]$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\right\|_2^2\leq&\frac{4\varepsilon^2}{\rho^2}\frac{\Delta^2}{\rho^2(\tau_k-\tau_{k+1})^2}\|\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j}\|_2^2\\&+\frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{A}^k}(\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j})\right\|_2^2+2(\varepsilon-1)^2\|\boldsymbol{A}_{\cdot,j}\|_2^2.\\&+\frac{2\Delta^2}{\rho^2(\tau_k-\tau_{k+1})^2}\left\|\boldsymbol{A}_{\cdot,j}^k\right\|_2^2+2\|\boldsymbol{E}_{\cdot,j}\|_2^2.\end{aligned}$$

Proof. First, we recall two conditions assumed in the Lemma that will be used in the proof: (1) $\|\boldsymbol{Z}-\rho\boldsymbol{A}\|\leq\Delta$ for some $\Delta\geq0$, (2) $\frac{1}{\varepsilon}\rho\leq\widehat{\rho}\leq\varepsilon\rho$ for some $\varepsilon\geq1$.

We will use notation $\lambda^*=s_k$, the kth singular value of $\boldsymbol{Z}$ for simplicity. We prove our Lemma in three steps.

Step 1. Fix a column index $j\in[p]$. Observe that

$$\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}=\left(\widehat{\boldsymbol{A}}_{\cdot,j}-\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})\right)+\left(\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})-\boldsymbol{A}_{\cdot,j}\right).$$

By choice, $\text{rank}(\widehat{\boldsymbol{A}})=k$. By definition (see (42)), we have that $\varphi_{\lambda^*}^{\boldsymbol{Z}}:\mathbb{R}^N\to\mathbb{R}^N$ is the projection operator onto the span of the top k left singular vectors of $\boldsymbol{Z}$, namely, $\text{span}\{u_1,\dots,u_k\}$. Therefore,

$$\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})-\boldsymbol{A}_{\cdot,j}\in\text{span}\{u_1,\dots,u_k\}^{\perp}$$

and by (43) (using Lemma I.1),

$$\widehat{\boldsymbol{A}}_{\cdot,j}-\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})=\frac{1}{\widehat{\rho}}\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{Z}_{\cdot,j})-\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})\in\text{span}\{u_1,\dots,u_k\}.$$

Hence, $\langle \widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}), \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) - \boldsymbol{A}_{\cdot,j} \rangle = 0$ and

$$\left\| \widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j} \right\|_2^2 = \left\| \widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) \right\|_2^2 + \left\| \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) - \boldsymbol{A}_{\cdot,j} \right\|_2^2 \tag{44}$$

by the Pythagorean theorem. It remains to bound the terms on the right hand side of (44).

Step 2. We begin by bounding the first term on the right hand side of (44). Again applying Lemma I.1, we can rewrite

$$\begin{aligned} \widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) &= \frac{1}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j}) - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) = \varphi^{\boldsymbol{Z}}_{\lambda^*}\left(\frac{1}{\widehat{\rho}} \boldsymbol{Z}_{\cdot,j} - \boldsymbol{A}_{\cdot,j} \right) \\ &= \frac{1}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}). \end{aligned}$$

Using the Parallelogram Law (or, equivalently, combining Cauchy-Schwartz and AM-GM inequalities), we obtain

$$\begin{aligned} \left\| \widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) \right\|_2^2 &= \left\| \frac{1}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) \right\|_2^2 \\ &\leq 2 \left\| \frac{1}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2 + 2 \left\| \frac{\rho - \widehat{\rho}}{\widehat{\rho}} \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) \right\|_2^2 \\ &\leq \frac{2}{\widehat{\rho}^2} \left\| \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2 + 2 \left(\frac{\rho - \widehat{\rho}}{\widehat{\rho}} \right)^2 \|\boldsymbol{A}_{\cdot,j}\|_2^2 \\ &\leq \frac{2\varepsilon^2}{\rho^2} \left\| \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2 + 2(\varepsilon - 1)^2 \|\boldsymbol{A}_{\cdot,j}\|_2^2. \end{aligned} \tag{45}$$

because Condition 2 implies $\frac{1}{\widehat{\rho}} \leq \frac{\varepsilon}{\rho}$ and $\left(\frac{\rho - \widehat{\rho}}{\widehat{\rho}} \right)^2 \leq (\varepsilon - 1)^2$.

Note that the first term of (45) can further be decomposed (using the Parallelogram Law and recalling $\boldsymbol{A} = \boldsymbol{A}^k + \boldsymbol{E}$, we have

$$\begin{aligned} &\left\| \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2 \\ &\quad \leq 2 \left\| \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) - \varphi^{\boldsymbol{A}^k}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2 + 2 \left\| \varphi^{\boldsymbol{A}^k}(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2. \end{aligned} \tag{46}$$

We now bound the first term on the right hand side of (46) separately. First, we apply the Davis-Kahan sinΘ Theorem (see [21, 47]) to arrive at the following inequality:

$$\|\mathcal{P}_{u_1,\dots,u_k} - \mathcal{P}_{\mu_1,\dots,\mu_k}\|_2 \le \frac{\|\boldsymbol{Z} - \rho\boldsymbol{A}\|}{\rho\tau_k - \rho\tau_{k+1}} \le \frac{\Delta}{\rho(\tau_k - \tau_{k+1})} \tag{47}$$

where $\mathcal{P}_{u_1,\dots,u_k}$ and $\mathcal{P}_{\mu_1,\dots,\mu_k}$ denote the projection operators onto the span of the top k left singular vectors of $\boldsymbol{Z}$ and $\boldsymbol{A}^k$, respectively. We utilized Condition 1 to bound $\|\boldsymbol{Z} - \rho\boldsymbol{A}\|_2 \le \Delta$. Then it follows that

$$\begin{aligned}\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}) - \varphi^{\boldsymbol{A}^k}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2 &\le \|\mathcal{P}_{u_1,\dots,u_k} - \mathcal{P}_{\mu_1,\dots,\mu_k}\|_2 \|\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}\|_2 \\ &\le \frac{\Delta}{\rho(\tau_k - \tau_{k+1})}\|\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}\|_2.\end{aligned}$$

Combining the inequalities together, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})\right\|_2^2 \le & \frac{4\varepsilon^2}{\rho^2}\frac{\Delta^2}{\rho^2(\tau_k - \tau_{k+1})^2}\|\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}\|_2^2 \\ & + \frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{A}^k}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2 + 2(\varepsilon - 1)^2\|\boldsymbol{A}_{\cdot,j}\|_2^2.\end{aligned} \tag{48}$$

Step 3. We now bound the second term of (44). Recalling $\boldsymbol{A} = \boldsymbol{A}^k + \boldsymbol{E}$ and using (47)

$$\begin{aligned}\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) - \boldsymbol{A}_{\cdot,j}\right\|_2^2 &= \left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}^k_{\cdot,j} + \boldsymbol{E}_{\cdot,j}) - \boldsymbol{A}^k_{\cdot,j} - \boldsymbol{E}_{\cdot,j}\right\|_2^2 \\ &\le 2\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}^k_{\cdot,j}) - \boldsymbol{A}^k_{\cdot,j}\right\|_2^2 + 2\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{E}_{\cdot,j}) - \boldsymbol{E}_{\cdot,j}\right\|_2^2 \\ &= 2\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}^k_{\cdot,j}) - \varphi^{\boldsymbol{A}^k}(\boldsymbol{A}^k_{\cdot,j})\right\|_2^2 + 2\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{E}_{\cdot,j}) - \boldsymbol{E}_{\cdot,j}\right\|_2^2 \\ &\le 2\|\mathcal{P}_{u_1,\dots,u_k} - \mathcal{P}_{\mu_1,\dots,\mu_k}\|^2\left\|\boldsymbol{A}^k_{\cdot,j}\right\|_2^2 + 2\|\boldsymbol{E}_{\cdot,j}\|_2^2 \\ &\le \frac{2\Delta^2}{\rho^2(\tau_k - \tau_{k+1})^2}\left\|\boldsymbol{A}^k_{\cdot,j}\right\|_2^2 + 2\|\boldsymbol{E}_{\cdot,j}\|_2^2.\end{aligned} \tag{49}$$

Inserting (48) and (49) back to (44) completes the proof. □

I.2 High probability events for conditioning

We define the following four events:

$$\begin{aligned}
\mathcal{E}_1 &:= \left\{ \|\boldsymbol{Z}-\rho\boldsymbol{A}\| \le \sqrt{C_1}\left(\sqrt{N}+\sqrt{p}\log^{\frac{3}{2}}(Np)\right)\right\}\\
\mathcal{E}_2 &:= \left\{ \left(1-\sqrt{\frac{20\log(Np)}{Np\rho}}\right)\rho \le \widehat{\rho} \le \frac{1}{1-\sqrt{\frac{20\log(Np)}{Np\rho}}}\rho \right\}\\
\mathcal{E}_3 &:= \left\{ \max_{j\in[p]} \left\| \boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j}\right\|_2^2 \le 11CK_\alpha^2 N\log^{\frac{2}{\alpha}}(Np)\right\}\\
\mathcal{E}_4 &:= \left\{ \max_{j\in[p]} \left\| \varphi^{\boldsymbol{A}^k}(\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j})\right\|_2^2 \le 11CK_\alpha^2 r\log^{\frac{2}{\alpha}}(Np)\right\}.
\end{aligned}$$

Here, $C_1 = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^2)$ for some constant $C>0$.

Observation 1: $\mathcal{E}_1$ occurs with high probability.

Lemma I.4. *Suppose that Properties 3.1, 3.3 for $\alpha \ge 1$ hold. Then, $\mathbb{P}(\mathcal{E}_1^c) \le \frac{2}{N^{10}p^{10}}$.*

Proof. The proof is complete by letting $\delta_1 = 10$ in Theorem H.1. □

Observation 2: $\mathcal{E}_2$ occurs with high probability.

Lemma I.5. *For any $\varepsilon > 1$,*

$$\mathbb{P}\left(\frac{1}{\varepsilon}\rho \le \widehat{\rho} \le \varepsilon\rho\right) \ge 1-2\exp\left(-\frac{(\varepsilon-1)^2}{2\varepsilon^2}Np\rho\right).$$

Proof. Recall that $\widehat{\rho} = \frac{1}{Np}\sum_{i=1}^N\sum_{j=1}^p \mathbb{1}(Z_{ij} \ne \star) \vee \frac{1}{Np}$. By the binomial Chernoff bound, for $\varepsilon > 1$,

$$\mathbb{P}(\widehat{\rho} > \varepsilon\rho) \le \exp\left(-\frac{(\varepsilon-1)^2}{\varepsilon+1}Np\rho\right), \quad \text{and}$$
$$\mathbb{P}\left(\widehat{\rho} < \frac{1}{\varepsilon}\rho\right) \le \exp\left(-\frac{(\varepsilon-1)^2}{2\varepsilon^2}Np\rho\right).$$

By the union bound,

$$\mathbb{P}\left(\frac{1}{\varepsilon}\rho \le \widehat{\rho} \le \varepsilon\rho\right) \ge 1-\mathbb{P}(\widehat{\rho} > \varepsilon\rho) - \mathbb{P}\left(\widehat{\rho} < \frac{1}{\varepsilon}\rho\right).$$

Noticing $\varepsilon+1 < 2\varepsilon < 2\varepsilon^2$ for all $\varepsilon > 1$ completes the proof. □

Remark I.2. *Let* $\varepsilon = \left(1 - \sqrt{\frac{20\log(Np)}{Np\rho}}\right)^{-1}$ *in Lemma I.5. Then,* $\mathbb{P}(\mathcal{E}_2^c) \le \frac{2}{N^{10}p^{10}}$.

Observation 3: $\mathcal{E}_3$ and $\mathcal{E}_4$ occur with high probability.

I.2.1 Two Helper Lemmas for $\mathcal{E}_3$ and $\mathcal{E}_4$

Lemma I.6. *Assume Properties 3.1, 3.3 hold. Then for any* $\alpha \ge 1$,

$$\|\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}\|_{\psi_\alpha} \le C(K_\alpha + 1), \qquad \forall j \in [p]$$

where $C > 0$ *is an absolute constant.*

Proof. Observe that

$$\begin{aligned}\|\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}\|_{\psi_\alpha} &= \sup_{u \in \mathbb{S}^{N-1}} \left\| u^T(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_{\psi_\alpha} \\ &= \sup_{u \in \mathbb{S}^{N-1}} \left\| u^T(\boldsymbol{Z} - \rho \boldsymbol{A}) e_j \right\|_{\psi_\alpha} \\ &= \sup_{u \in \mathbb{S}^{N-1}} \left\| \sum_{i=1}^n u_i (\boldsymbol{Z}_{i,\cdot} - \rho \boldsymbol{A}_{i,\cdot}) e_j \right\|_{\psi_\alpha} \\ &\overset{(a)}{\le} C \sup_{u \in \mathbb{S}^{N-1}} \left(\sum_{i=1}^n u_i^2 \|(\boldsymbol{Z}_{i,\cdot} - \rho \boldsymbol{A}_{i,\cdot}) e_j\|_{\psi_\alpha}^2 \right)^{1/2} \\ &\le C \max_{i \in [N]} \|\boldsymbol{Z}_{i,\cdot} - \rho \boldsymbol{A}_{i,\cdot}\|_{\psi_\alpha},\end{aligned}$$

where (a) follows from Lemma H.4. Then the conclusion follows from Lemma H.5. □

Lemma I.7. *Let* $W_1, ..., W_n$ *be a sequence of* ψ_α*-random variables for some* $\alpha \ge 1$. *For any* $t \ge 0$,

$$\mathbb{P}\left(\sum_{i=1}^n W_i^2 > t \right) \le 2 \sum_{i=1}^n \exp\left(-\left(\frac{t}{n \|W_i\|_{\psi_\alpha}^2} \right)^{\alpha/2} \right).$$

Proof. Note that $\sum_{i=1}^n W_i^2 > t$ implies that there exists at least one $i \in [n]$ with $W_i^2 > \frac{t}{n}$. By the union bound,

$$\mathbb{P}\left(\sum_{i=1}^n W_i^2 > t \right) \le \sum_{i=1}^n \mathbb{P}\left(W_i^2 > \frac{t}{n} \right) \qquad \le \sum_{i=1}^n \mathbb{P}\left(|W_i| > \sqrt{\frac{t}{n}} \right) \le \sum_{i=1}^n 2\exp\left(-\left(\frac{t}{n \|W_i\|_{\psi_\alpha}^2} \right)^{\alpha/2} \right).$$

□

Lemma I.8. *Suppose Properties 3.1, 3.3 hold. Then,*

$$\mathbb{P}(\mathcal{E}_3^c) \le \frac{2}{N^{10}p^{10}}.$$

Proof. Fix $j \in [p]$. Let $e_i \in \mathbb{R}^N$ denote the i-th canonical basis of $\mathbb{R}^N$ (column vector representation). Note that

$$\left\| \boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j} \right\|_2^2 = \sum_{i=1}^{N} \left(e_i^T (\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right)^2$$

and $e_i^T(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j})$ is a ψ_α-random variable with $\left\| e_i^T(\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_{\psi_\alpha} \le \| \boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j} \|_{\psi_\alpha}$. By Lemma I.6, $\| \boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j} \|_{\psi_\alpha} \le C(K_\alpha + 1)$ for all $j \in [p]$. By Lemma I.7 and the union bound,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_3^c) &\le \sum_{j=1}^{p} \mathbb{P}\left(\left\| \boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j} \right\|_2^2 > 11 C^2 (K_\alpha + 1)^2 N \log^{\frac{2}{\alpha}}(Np) \right) \\
&\le 2 \sum_{j=1}^{p} \sum_{i=1}^{N} \exp(-11 \log(Np)) \\
&= \frac{2}{N^{10}p^{10}}.
\end{aligned}$$

□

Lemma I.9. *Suppose properties 3.1, 3.3 hold. Then,*

$$\mathbb{P}(\mathcal{E}_4^c) \le \frac{2}{N^{10}p^{10}}.$$

Proof. Recall that $\text{rank}(\boldsymbol{A}^k) = k$. We write

$$\left\| \varphi^{\boldsymbol{A}^k} (\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right\|_2^2 = \sum_{i=1}^{k} \left(u_i^T (\boldsymbol{Z}_{\cdot,j} - \rho \boldsymbol{A}_{\cdot,j}) \right)^2,$$

where $u_1, \ldots, u_k$ denote the left singular vectors of $\boldsymbol{A}^k$. The proof has the same structure with that of Lemma I.8 with $u_1, \ldots, u_k$ in place of $e_1, \ldots, e_n$. □

I.3 Completing Proof of Lemma 3.1

Proof of Lemma 3.1. Recall that our goal is to establish

$$\mathbb{E}[\|\boldsymbol{Z}^{\text{HSVT},k}-\boldsymbol{A}\|_{2,\infty}^2]\leq\frac{C(K_\alpha^2+1)}{\rho^2}\Big(k+\frac{N\Delta^2}{\rho^2(\tau_k-\tau_{k+1})^2}\Big)\log^{\frac{2}{\alpha}}Np+2\|\boldsymbol{A}^k-\boldsymbol{A}\|_{2,\infty}^2,$$

where $C>0$ is a universal constant. To that end, define $E\triangleq\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$. By Lemmas I.4, I.5, I.8 and I.9, it follows that

$$\mathbb{P}(E^c)\leq\mathbb{P}(\mathcal{E}_1^c\cup\mathcal{E}_2^c\cup\mathcal{E}_3^c\cup\mathcal{E}_4^c)\ \leq\ \frac{8}{N^{10}p^{10}}.$$

Observe (with $\widehat{\boldsymbol{A}}=\boldsymbol{Z}^{\text{HSVT},k}$),

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}^2]&=\mathbb{E}\max_{j\in[p]}\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\Big\|_2^2\\&=\mathbb{E}\Big[\max_{j\in[p]}\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\Big\|_2^2\cdot\mathbb{1}(E)\Big]+\mathbb{E}\Big[\max_{j\in[p]}\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\Big\|_2^2\cdot\mathbb{1}(E^c)\Big].\end{aligned}\tag{50}$$

In the rest of the proof, we upper bound the two terms in (50) separately.

Upper bound on the first term in (50). Under event E, from Lemma I.3, we have

$$\max_{j\in[p]}\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\Big\|_2^2\leq\frac{C(K_\alpha+1)^2}{\rho^2}\Big(\frac{\Delta^2N}{\rho^2(\tau_r-\tau_{r+1})^2}+r\Big)\log^{\frac{2}{\alpha}}(Np)+2\max_{j\in[p]}\|\boldsymbol{E}_{\cdot,j}\|_2^2.$$

where $C>0$ is an absolute constant. To see this, note that $\varepsilon^2\leq10$ since $\rho\geq\frac{64\log(Np)}{Np}$; $\|\boldsymbol{A}_j^k\|_2^2\leq\|\boldsymbol{A}_j\|_2^2\leq N$, again appealing to the contraction property of the HSVT operator (refer to Lemma I.2 and Property 3.1). Since $\mathbb{P}(E)\leq1$, it follows that

$$\mathbb{E}\Big[\Big\|\widehat{\boldsymbol{A}}-\boldsymbol{A}\Big\|_{2,\infty}^2\cdot\mathbb{1}(E)\Big]\leq\frac{C(K_\alpha+1)^2}{\rho^2}\Big(\frac{\Delta^2N}{\rho^2(\tau_r-\tau_{r+1})^2}+r\Big)\log^{\frac{2}{\alpha}}(Np)+2\max_{j\in[p]}\|\boldsymbol{E}_{\cdot,j}\|_2^2.\tag{51}$$

Upper bound on the second term in (50). To begin with, we note that for any $j\in[p]$,

$$\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\Big\|_2\leq\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}\Big\|_2+\|\boldsymbol{A}_{\cdot,j}\|_2$$

by triangle inequality. By the model assumption, the covariates are bounded (Property 3.1) and $\|\boldsymbol{A}_{\cdot,j}\|_2 \le \sqrt{N}$ for all $j \in [p]$. By definition, for any $j \in [p]$,

$$\widehat{\boldsymbol{A}}_{\cdot,j} = \frac{1}{\widehat{\rho}} \text{HSVT}_\lambda(\boldsymbol{Z})_{\cdot,j}$$

for a given threshold $\lambda = s_k$, the kth singular value of $\boldsymbol{Z}$. Therefore,

$$\|\widehat{\boldsymbol{A}}_{\cdot,j}\|_2 = \frac{1}{\widehat{\rho}} \|\text{HSVT}_\lambda(\boldsymbol{Z})_{\cdot,j}\|_2 \overset{(a)}{\le} Np \|\text{HSVT}_\lambda(\boldsymbol{Z})_{\cdot,j}\|_2 \overset{(b)}{\le} Np \|\boldsymbol{Z}_{\cdot,j}\|_2.$$

Here, (a) follows from $\widehat{\rho} \ge \frac{1}{Np}$; and (b) follows from Lemma I.2 – the HSVT operator is a contraction on the columns.

$$\begin{aligned}
\max_{j\in[p]} \|\widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\|_2 &\le \max_{j\in[p]} \|\widehat{\boldsymbol{A}}_{\cdot,j}\|_2 + \max_{j\in[p]} \|\boldsymbol{A}_{\cdot,j}\|_2 \\
&\le Np \max_{j\in[p]} \|\boldsymbol{Z}_{\cdot,j}\|_2 + \sqrt{N} \\
&\le (N^{\frac{3}{2}} p + \sqrt{N}) + N^{\frac{3}{2}} p \max_{ij} |\eta_{ij}| \\
&\le 2N^{\frac{3}{2}} p \Big(1 + \max_{ij} |\eta_{ij}|\Big) \qquad (52)
\end{aligned}$$

because $\max_{j\in[p]} \|\boldsymbol{Z}_{\cdot,j}\|_2 \le \sqrt{N} \max_{i,j} |Z_{ij}| \le \sqrt{N} \max_{i,j} |A_{ij} + \eta_{ij}| \le \sqrt{N}(1 + \max_{i,j} |\eta_{ij}|)$. Now we apply Cauchy-Schwarz inequality on $\mathbb{E}[\max_{j\in[p]} \|\widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\|_2^2 \cdot \mathbb{1}(E^c)]$ to obtain

$$\begin{aligned}
\mathbb{E}\Big[\max_{j\in[p]} \|\widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\|_2^2 \cdot \mathbb{1}(E^c)\Big] &\le \mathbb{E}\Big[\max_{j\in[p]} \|\widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\|_2^4\Big]^{\frac{1}{2}} \cdot \mathbb{E}\Big[\mathbb{1}(E^c)\Big]^{\frac{1}{2}} \\
&= \mathbb{E}\Big[\max_{j\in[p]} \|\widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\|_2^4\Big]^{\frac{1}{2}} \cdot \mathbb{P}(E^c)^{\frac{1}{2}} \\
&\overset{(a)}{\le} 4N^3 p^2 \mathbb{E}\Big[\Big(1 + \max_{ij} |\eta_{ij}|\Big)^4\Big]^{\frac{1}{2}} \cdot \mathbb{P}(E^c)^{\frac{1}{2}} \\
&\overset{(b)}{\le} 8\sqrt{2} N^3 p^2 \Big(1 + \mathbb{E}[\max_{ij} |\eta_{ij}|^4]\Big)^{\frac{1}{2}} \cdot \mathbb{P}(E^c)^{\frac{1}{2}} \\
&\overset{(c)}{\le} 8\sqrt{2} N^3 p^2 \Big(1 + \mathbb{E}[\max_{ij} |\eta_{ij}|^4]^{\frac{1}{2}}\Big) \cdot \mathbb{P}(E^c)^{\frac{1}{2}}. \qquad (53)
\end{aligned}$$

Here, (a) follows from (52); and (b) follows from Jensen's inequality:

$$\mathbb{E}\Big[\Big(1 + \max_{ij} |\eta_{ij}|\Big)^4\Big] = \mathbb{E}\Big[\Big(\frac{1}{2}(2 + 2\max_{ij} |\eta_{ij}|)\Big)^4\Big] \le \mathbb{E}\Big[\frac{1}{2}\Big(2^4 + (2\max_{ij} |\eta_{ij}|)^4\Big)\Big]$$

$$=8\mathbb{E}\Big[1+\max_{ij}|\eta_{ij}|^4\Big]=8\Big(1+\mathbb{E}[\max_{ij}|\eta_{ij}|^4]\Big);$$

and (c) follows from the trivial inequality: $\sqrt{A+B}\leq\sqrt{A}+\sqrt{B}$ for any $A,B\geq 0$.

Now it remains to find an upper bound for $\mathbb{E}[\max_{ij}|\eta_{ij}|^4]$. Note that for any $\alpha>0$ and $\theta\geq 1$, η_{ij} being a ψ_α-random variable implies that $|\eta_{ij}|^\theta$ is a $\psi_{\alpha/\theta}$-random variable. With the choice of $\theta=4$, we have that

$$\mathbb{E}\max_{ij}|\eta_{ij}|^4\leq C_1K_\alpha^4\log^{\frac{4}{\alpha}}(Np) \tag{54}$$

for some absolute constant $C_1>0$ by Lemma D.5 (also see Remark D.1). Inserting (54) to (53) yields

$$\begin{aligned}\mathbb{E}\Big[\max_{j\in[p]}\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\|_2^2\cdot\mathbb{1}(\mathcal{E}^c)\Big]&\leq 8\sqrt{2}N^3p^2\Big(1+{C_1'}^{1/2}K_\alpha^2\log^{\frac{2}{\alpha}}(Np)\Big)\cdot\mathbb{P}(E^c)^{\frac{1}{2}}\\&\overset{(a)}{\leq}32\Big(1+{C_1}^{1/2}K_\alpha^2\log^{\frac{2}{\alpha}}(Np)\Big)\frac{1}{N^2p^2},\end{aligned} \tag{55}$$

where (a) follows from recalling that $\mathbb{P}(E^c)\leq 8/N^{10}p^{10}$.

Concluding the Proof. Thus, combining (51) and (55) in (50) and noticing that term in (55) is smaller order term than that in (51), by defining appropriate constant $C>0$, we obtain:

$$\begin{aligned}\mathbb{E}[\|\widehat{\boldsymbol{A}}-\boldsymbol{A}\|_{2,\infty}^2]\leq&\frac{C(K_\alpha+1)^2}{\rho^2}\Big(\frac{\Delta^2N}{\rho^2(\tau_r-\tau_{r+1})^2}+r\Big)\log^{\frac{2}{\alpha}}(Np)+2\max_{j\in[p]}\|\boldsymbol{E}_{\cdot,j}\|_2^2\\&+\frac{C}{N^2p^2}\Big(1+K_\alpha^2\log^{\frac{2}{\alpha}}(Np)\Big),\end{aligned}$$

with $\Delta=\sqrt{C^*}\Big(\sqrt{N}+\sqrt{p}\log^{\frac{3}{2}}(Np)\Big)$ and $C^*=C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^2)$.

The proof is complete by defining $C'=C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ and simplifying the bound further in a straightforward manner. □

J Proof of Lemma 3.2

The proof of Lemma 3.2 follows very closely the structure of the proof of Lemma 3.1. The key difference is Lemma I.3 no longer holds as is, and needs to be redefined for $\boldsymbol{A}^{(\mathrm{lr})}$ instead of $\boldsymbol{A}^k$.

Lemma J.1. *Suppose that (1) $\|\boldsymbol{Z}-\rho\boldsymbol{A}\|\leq\Delta$ for some $\Delta\geq 0$ and (2) $\frac{1}{\varepsilon}\rho\leq\widehat{\rho}\leq\varepsilon\rho$ for some $\varepsilon\geq 1$.*

Let $\boldsymbol{A}=\boldsymbol{A}^{(\mathrm{lr})}+\boldsymbol{E}^{(\mathrm{lr})}$. Let $r=rank(\boldsymbol{A}^{(\mathrm{lr})})$ and τ_r denote the r-th singular value of $\boldsymbol{A}^{(\mathrm{lr})}$. Let $\widehat{\boldsymbol{A}}=\boldsymbol{Z}^{\mathrm{HSVT},r}$. Then for any $j\in[p]$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}\right\|_2^2\leq\frac{8\varepsilon^2}{\rho^4}\Big(\frac{\Delta^2}{\tau_r^2}+\frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{\tau_r^2}\Big)\Big(\|\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j}\|_2^2+\left\|\boldsymbol{A}_{\cdot,j}^{(\mathrm{lr})}\right\|_2^2\Big)\\+\frac{4\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{A}^{(\mathrm{lr})}}(\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j})\right\|_2^2+2(\varepsilon-1)^2\|\boldsymbol{A}_{\cdot,j}\|_2^2\\+2\left\|\boldsymbol{E}_{\cdot,j}^{(\mathrm{lr})}\right\|_2^2.\end{aligned}$$

The proof of Lemma J.1 is almost identical to the proof of Lemma I.3, except the replacement of the subspace perturbation bound (47) with a new one in (58). Roughly speaking, we control the principal angle between the top-r left singular space of $\boldsymbol{Z}$ and the column space of $\boldsymbol{A}^{(\mathrm{lr})}$ by means of triangle inequality, using the column space of $\boldsymbol{A}^k$ as an intermeidary. Despite the similarity to the proof of Lemma I.3, we present the full proof of Lemma J.1 for future reference in synthetic control literature.

Proof. We will use notation $\lambda^*=s_r$, the rth singular value of $\boldsymbol{Z}$ for simplicity. We prove our Lemma in three steps.

Step 1. Fix a column index $j\in[p]$. Observe that

$$\widehat{\boldsymbol{A}}_{\cdot,j}-\boldsymbol{A}_{\cdot,j}=\Big(\widehat{\boldsymbol{A}}_{\cdot,j}-\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})\Big)+\Big(\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})-\boldsymbol{A}_{\cdot,j}\Big).$$

By choice, $\mathrm{rank}(\widehat{\boldsymbol{A}})=r$. By definition (see (42)), we have that $\varphi_{\lambda^*}^{\boldsymbol{Z}}:\mathbb{R}^N\to\mathbb{R}^N$ is the projection operator onto the span of the top r left singular vectors of $\boldsymbol{Z}$, namely, $\mathrm{span}\{u_1,...,u_r\}$. Therefore,

$$\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})-\boldsymbol{A}_{\cdot,j}\in\mathrm{span}\{u_1,...,u_r\}^{\perp}$$

and by (43) (using Lemma I.1),

$$\widehat{\boldsymbol{A}}_{\cdot,j}-\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})=\frac{1}{\widehat{\rho}}\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{Z}_{\cdot,j})-\varphi_{\lambda^*}^{\boldsymbol{Z}}(\boldsymbol{A}_{\cdot,j})\in\mathrm{span}\{u_1,...,u_r\}.$$

Hence, $\langle \widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}), \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) - \boldsymbol{A}_{\cdot,j}\rangle = 0$ and

$$\left\|\widehat{\boldsymbol{A}}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\right\|_2^2 = \left\|\widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})\right\|_2^2 + \left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) - \boldsymbol{A}_{\cdot,j}\right\|_2^2 \tag{56}$$

by the Pythagorean theorem. It remains to bound the terms on the right hand side of (56).

Step 2. We begin by bounding the first term on the right hand side of (56). Again applying Lemma I.1, we can rewrite

$$\begin{aligned}\widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) &= \frac{1}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j}) - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}) = \varphi^{\boldsymbol{Z}}_{\lambda^*}\Big(\frac{1}{\widehat{\rho}}\boldsymbol{Z}_{\cdot,j} - \boldsymbol{A}_{\cdot,j}\Big)\\ &= \frac{1}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j}).\end{aligned}$$

Using the Parallelogram Law (or, equivalently, combining Cauchy-Schwartz and AM-GM inequalities), we obtain

$$\begin{aligned}\left\|\widehat{\boldsymbol{A}}_{\cdot,j} - \varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})\right\|_2^2 &= \left\|\frac{1}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}) + \frac{\rho - \widehat{\rho}}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})\right\|_2^2\\ &\le 2\left\|\frac{1}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2 + 2\left\|\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})\right\|_2^2\\ &\le \frac{2}{\widehat{\rho}^2}\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2 + 2\Big(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\Big)^2\|\boldsymbol{A}_{\cdot,j}\|_2^2\\ &\le \frac{2\varepsilon^2}{\rho^2}\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2 + 2(\varepsilon - 1)^2\|\boldsymbol{A}_{\cdot,j}\|_2^2.\end{aligned} \tag{57}$$

because Condition 2 implies $\frac{1}{\widehat{\rho}} \le \frac{\varepsilon}{\rho}$ and $\left(\frac{\rho - \widehat{\rho}}{\widehat{\rho}}\right)^2 \le (\varepsilon - 1)^2$.

Note that the first term of (57) can further be decomposed (using the Parallelogram Law and recalling $\boldsymbol{A} = \boldsymbol{A}^{(\mathrm{lr})} + \boldsymbol{E}^{(\mathrm{lr})}$, we have

$$\begin{aligned}&\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2\\ &\quad\le 2\left\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j}) - \varphi^{\boldsymbol{A}^{(\mathrm{lr})}}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2 + 2\left\|\varphi^{\boldsymbol{A}^{(\mathrm{lr})}}(\boldsymbol{Z}_{\cdot,j} - \rho\boldsymbol{A}_{\cdot,j})\right\|_2^2.\end{aligned}$$

We now bound the first term on the right hand side of (46) separately. First, we apply the Davis-Kahan sinΘ Theorem (see [21, 47]) to arrive at the following inequality:

$$\begin{aligned}\|\mathcal{P}_{u_1,\dots,u_r}-\mathcal{P}_{\mu_1,\dots,\mu_r}\|_2 &\le \frac{\|\boldsymbol{Z}-\rho\boldsymbol{A}^{(\mathrm{lr})}\|_2}{\rho\tau_r} \\ &\le \frac{\|\boldsymbol{Z}-\rho\boldsymbol{A}\|_2}{\rho\tau_r}+\frac{\|\rho\boldsymbol{A}-\rho\boldsymbol{A}^{(\mathrm{lr})}\|_2}{\rho\tau_r} \\ &\le \frac{\Delta}{\rho\tau_r}+\frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_2}{\rho\tau_r},\end{aligned} \tag{58}$$

where $\mathcal{P}_{u_1,\dots,u_r}$ and $\mathcal{P}_{\mu_1,\dots,\mu_r}$ denote the projection operators onto the span of the top r left singular vectors of $\boldsymbol{Z}$ and $\boldsymbol{A}^{(\mathrm{lr})}$, respectively. We utilized Condition 1 to bound $\|\boldsymbol{Z}-\rho\boldsymbol{A}\|_2\le\Delta$. Then it follows that

$$\begin{aligned}\Big\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j})-\varphi^{\boldsymbol{A}^{(\mathrm{lr})}}(\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j})\Big\|_2 &\le \|\mathcal{P}_{u_1,\dots,u_r}-\mathcal{P}_{\mu_1,\dots,\mu_r}\|_2\|\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j}\|_2 \\ &\le \Big(\frac{\Delta}{\rho\tau_r}+\frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_2}{\rho\tau_r}\Big)\|\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j}\|_2.\end{aligned}$$

Combining the inequalities together, we have

$$\begin{aligned}\Big\|\widehat{\boldsymbol{A}}_{\cdot,j}-\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})\Big\|_2^2 \le& \frac{8\varepsilon^2}{\rho^4}\Big(\frac{\Delta^2}{\tau_r^2}+\frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{\tau_r^2}\Big)\|\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j}\|_2^2 \\ &+\frac{4\varepsilon^2}{\rho^2}\Big\|\varphi^{\boldsymbol{A}^{(\mathrm{lr})}}(\boldsymbol{Z}_{\cdot,j}-\rho\boldsymbol{A}_{\cdot,j})\Big\|_2^2+2(\varepsilon-1)^2\|\boldsymbol{A}_{\cdot,j}\|_2^2.\end{aligned} \tag{59}$$

Step 3. We now bound the second term of (56). Recalling $\boldsymbol{A}=\boldsymbol{A}^{(\mathrm{lr})}+\boldsymbol{E}^{(\mathrm{lr})}$ and using (58)

$$\begin{aligned}\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}_{\cdot,j})-\boldsymbol{A}_{\cdot,j}\|_2^2 &= \Big\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j}+\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j})-\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j}-\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2 \\ &\le 2\Big\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j})-\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2+2\Big\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j})-\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2 \\ &= 2\Big\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j})-\varphi^{\boldsymbol{A}^{(\mathrm{lr})}}(\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j})\Big\|_2^2+2\Big\|\varphi^{\boldsymbol{Z}}_{\lambda^*}(\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j})-\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2 \\ &\le 2\|\mathcal{P}_{u_1,\dots,u_r}-\mathcal{P}_{\mu_1,\dots,\mu_r}\|^2\Big\|\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2+2\Big\|\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2 \\ &\le 4\Big(\frac{\Delta^2}{\rho^2\tau_r^2}+\frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{\rho^2\tau_r^2}\Big)\Big\|\boldsymbol{A}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2+2\Big\|\boldsymbol{E}^{(\mathrm{lr})}_{\cdot,j}\Big\|_2^2.\end{aligned} \tag{60}$$

Inserting (59) and (60) back to (56) completes the proof. □

J.1 Completing Proof of Lemma 3.2

Proof of Lemma 3.2. Proof follows in an identical fashion to that of Lemma 3.1 (see Section I.3) and using the bound in Lemma J.1 instead of the one in Lemma I.3. □

K Proof of Corollary 3.4

Proof. From Proposition 3.3, we have that $r \le C(\zeta,K)\delta^{-K}$ and $\|\boldsymbol{E}^{(\mathrm{lr})}\|_\infty \le \mathcal{L}\cdot\delta^\zeta$. So, $\tau_r^2 \ge CNp/r \ge CNp/(C(\zeta,K)\delta^{-K})$.

$$
\begin{aligned}
MSE_\Omega(\widehat{Y}) &\le \frac{4\sigma^2 r}{n} + \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(\frac{n\vee p\vee\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{\tau_r^2} + \frac{r}{n}\right)\log^5(np) + \frac{6\|\beta^*\|_1^2}{n}\|\boldsymbol{E}^{(\mathrm{lr})}\|_{2,\infty}^2 + \frac{20}{n}\|\phi\|_2^2 \\
&\le \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(\frac{n\vee p\vee\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{\tau_r^2} + \frac{r}{n} + \frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_{2,\infty}^2}{n}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2 \\
&\le \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(\frac{r\|\boldsymbol{E}^{(\mathrm{lr})}\|_2^2}{Np} + \frac{r}{n\wedge p} + \frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_{2,\infty}^2}{n}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2 \\
&\le \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(r\|\boldsymbol{E}^{(\mathrm{lr})}\|_\infty^2 + \frac{r}{n\wedge p} + \frac{\|\boldsymbol{E}^{(\mathrm{lr})}\|_{2,\infty}^2}{n}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2 \\
&\le \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(r\|\boldsymbol{E}^{(\mathrm{lr})}\|_\infty^2 + \frac{r}{n\wedge p}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2 \\
&\le \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(C(\zeta,K)\delta^{-K}\mathcal{L}^2\cdot\delta^{2\zeta} + \frac{C(\zeta,K)\delta^{-K}}{n\wedge p}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2 \\
&\le \frac{C'C(\zeta,K)\mathcal{L}^2\|\beta^*\|_1^2}{\rho^4}\left(\delta^{-K}\cdot\delta^{2\zeta} + \frac{\delta^{-K}}{n\wedge p}\right)\log^5(np) + \frac{20}{n}\|\phi\|_2^2
\end{aligned}
$$

Substituting $\delta = (1/(n\wedge p))^{1/2\zeta}$ completes the proof. □

L Proof of Proposition 3.4

Proof. We have

$$
M = \sum_{i=1}^{p} v_i \boldsymbol{X}_{\cdot,i}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} v_i \boldsymbol{X}_{\cdot,i} + \sum_{j=k+1}^{p} v_j \boldsymbol{X}_{\cdot,j} \\
&= \sum_{i=1}^{k} v_i \boldsymbol{X}_{\cdot,i} + \sum_{j=k+1}^{p} v_j \Big(\sum_{i=1}^{k} c_i(j) \boldsymbol{X}_{\cdot,i} \Big) \\
&= \sum_{i=1}^{k} v_i \boldsymbol{X}_{\cdot,i} + \sum_{i=1}^{k} \boldsymbol{X}_{\cdot,i} \Big(\sum_{j=k+1}^{p} c_i(j) v_j \Big) \\
&= \sum_{i=1}^{k} \Big(v_i + \sum_{j=k+1}^{p} c_i(j) v_j \Big) \boldsymbol{X}_{\cdot,i}.
\end{aligned}
$$

Define $v_i^* = v_i + \sum_{j=k+1}^{p} c_i(j) v_j$ for $i \in [k]$ and 0 for $i \notin [k]$. Then $\|v^*\|_0 \leq k$. Further,

$$
\|v^*\|_1 = \sum_{i=1}^{k} \left| \Big(v_i + \sum_{j=k+1}^{p} c_i(j) v_j \Big) \right| \leq C'' \sum_{i=1}^{k} \Big(|v_i| + \sum_{j=k+1}^{p} |v_j| \Big) \leq C'' k \|v\|_1
$$

□

M Proof of Theorem 3.2

The proof of Theorem 3.2 follows the standard approach in terms of establishing generalization error bounds using Rademacher complexity (see [10] and references therein). We note two important contributions: (1) relating our notion of generalization error to the standard definitions; (2) arguing that the Rademacher complexity of our matrix estimation regression algorithm (using HSVT) can be identified with the Rademacher complexity of regression with ℓ_0-regularization.

M.1 Background

Notation, Setup. We consider PCR with parameter k for some $k \geq 1$. Recall that the training sample set $\Omega \subset [N]$, with $|\Omega| = n$, is sampled uniformly at random and without replacement from $[N]$. Further, as argued in Proposition 2.1, PCR with parameter k is equivalent to Linear Regression with pre-processing of noisy covariates using HSVT. Hence, we let $\widehat{\boldsymbol{A}} = \boldsymbol{Z}^{\text{HSVT},k}$ and $\widehat{\beta} = \beta^{\text{HSVT},k}$.

Generalization error and Rademacher complexity. We measure the quality of our estimates through the following two quantities of error. For any hypothesis $\beta \in \mathbb{R}^p$ and training set Ω, the empirical error is

$$\widehat{\mathcal{E}}_\Omega(\beta) = \frac{1}{n}\sum_{\omega \in \Omega}\Big(\widehat{\boldsymbol{A}}_{\omega,\cdot}\beta - \boldsymbol{A}_{\omega,\cdot}\beta^*\Big)^2. \tag{61}$$

Similarly, we define the overall error as

$$\mathcal{E}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\Big(\widehat{\boldsymbol{A}}_{i,\cdot}\beta - \boldsymbol{A}_{i,\cdot}\beta^*\Big)^2. \tag{62}$$

For any linear hypothesis class $\mathcal{F} \subset \mathbb{R}^p$, define the generalization error as the supremum of the gap between (61) and (62) over $\mathcal{F}$. Precisely, for a given training set Ω,

$$\phi(\Omega) = \sup_{\beta \in \mathcal{F}}\Big(\mathcal{E}(\beta) - \widehat{\mathcal{E}}_\Omega(\beta)\Big). \tag{63}$$

Next, we define the notion of Rademacher complexity, wich has been very effective to bound the generalization error. To begin with, the Rademacher complexity of a set $A \subset \mathbb{R}^n$ is defined as

$$R(A) = \mathbb{E}_\sigma\left[\sup_{a \in A}\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right],$$

where $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher variables, which are uniformly distributed on $\{-1,1\}$, and the expectation above is taken with respect to their randomness. This has been naturally extended for the setting of prediction problems as follows: given a collection of real-valued response variables and covariates, say (Y_i,X_i), $i \in [n]$, a collection of real-valued functions or hypotheses $\mathcal{G}$ that map covariates to real values, and loss function $L:\mathbb{R}^2 \to [0,\infty)$ that measures the error or loss in prediction for a given function, define

$$R_S(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g(X_i)\right],\ \ R_S(L \circ \mathcal{G}) \ = \mathbb{E}_\sigma\left[\sup_{g \in \mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i L(Y_i, g(X_i))\right].$$

In our setting, the covariates that the predictor uses are the denoised rows of $\widehat{\boldsymbol{A}}$, denoted as$\{\widehat{\boldsymbol{A}}_{1,\cdot},\ldots,\widehat{\boldsymbol{A}}_{N,\cdot}\}$. The loss function of interest is the quadratic function: $\ell(y,y') = (y-y')^2$. The ideal response variable of our interest is $\boldsymbol{A}_{i,\cdot}\beta^*$ for $i\in[N]$. Given that our algorithm observes (noisy) response variables in the index set Ω, we shall use the sample set $\{(\boldsymbol{A}_{\omega,\cdot}\beta^*,\widehat{\boldsymbol{A}}_{\omega,\cdot}):\omega\in\Omega\}$.

It turns out that the appropriate adaptation of the Rademacher complexity for our setting is as follows: Let $\mathcal{D}$ denote the distribution of the observations Z_{ij} (i.e., the randomness in the measurements). Hence, $\widehat{\boldsymbol{A}}$ is a random matrix as it derived from $\boldsymbol{Z}$. Then,

$$R_n(\mathcal{F})=\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\left(\frac{1}{n}\sum_{\omega\in\Omega}\sigma_\omega\widehat{\boldsymbol{A}}_{\omega,\cdot}\beta\right)\right]$$
$$R_n(\ell\circ\mathcal{F})=\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\left(\frac{1}{n}\sum_{\omega\in\Omega}\sigma_\omega\ell(\boldsymbol{A}_{\omega,\cdot}\beta^*,\widehat{\boldsymbol{A}}_{\omega,\cdot}\beta)\right)\right],$$

where $\mathbb{E}_\Omega$ is taken with respect to selecting $\Omega\subset[N]$ uniformly at random from $[N]$ without replacement (with $|\Omega|=n$).

Rademacher Class - Sparse Linear Models. Define $\mathcal{F}_{(a,b)}\subset\mathbb{R}^p$ for $a\in\mathbb{N},b\in\mathbb{R}$ as

$$\mathcal{F}_{(a,b)}:=\{\beta\in\mathbb{R}^p:\|\beta\|_0\leq a,\|\beta\|_1\leq b\}.$$

We denote $\mathcal{F}_{(\cdot,b)}$ as the case where there is no restriction on a, i.e., $\beta\in\mathcal{F}_{(\cdot,b)}$ has no constraint in its $\|\cdot\|_0$-norm. We then have the following proposition,

Proposition M.1. *Assume $\widehat{\boldsymbol{A}}$ satisfies* (13) *in Proposition 3.4. Then,*

$$R_n(\mathcal{F}_{(\cdot,\ \|\widehat{\beta}\|_1)})\leq R_n(\mathcal{F}_{(k,\ C''k\|\widehat{\beta}\|_1)}),\quad R_n(\ell\circ\mathcal{F}_{(\cdot,\ \|\widehat{\beta}\|_1})\leq R_n(\ell\circ\mathcal{F}_{(k,\ C''k\|\widehat{\beta}\|_1)})$$

where C'' is defined as in Proposition 3.4.

Proof. By definition, $\widehat{\boldsymbol{A}}$ has rank k. Then by Proposition 3.4 for $\widehat{\beta}$, there exists an k-sparse vector $\beta'\in\mathbb{R}^p$ such that

$$\widehat{\boldsymbol{A}}\cdot\widehat{\beta}=\widehat{\boldsymbol{A}}\cdot\beta',\text{ s.t. }\|\beta'\|_1\leq C''\|\widehat{\beta}\|_1.$$

Observe that due to the equality, we have,

$$\widehat{\mathcal{E}}_{\Omega}(\widehat{\beta}) = \widehat{\mathcal{E}}_{\Omega}(\beta^{'})$$

$$\mathcal{E}_{\Omega}(\widehat{\beta}) = \mathcal{E}_{\Omega}(\beta^{'}).$$

Appealing to the definitions of $R_n(\cdot)$ and $R_n(\ell \circ \cdot)$ completes the proof. □

For the remainder of Section M, we define $B := C'' \cdot k \cdot \|\widehat{\beta}\|_1$ and overload notation and define $\mathcal{F} := \mathcal{F}_{(k,B)}$.

M.2 Helper Lemmas M.1 and M.5 to Prove Theorem 3.2

M.2.1 Lemma M.1

Lemma M.1. *Let $\phi(\Omega)$ be defined as in* (63). *Let Ω be random subset of $[N]$ of size n that is chosen uniformly at random without replacement. Then,*

$$\mathbb{E}_{\Omega|\mathcal{D}}[\phi(\Omega)] \le 2R_n(\ell \circ \mathcal{F}).$$

Proof. Let $\Omega = \{i_1, ..., i_n\}$. Further, let $\Omega' = \{i'_1, ..., i'_n\}$ be a "ghost sample", i.e., Ω' is an independent set of n locations sampled uniformly at random and without replacement from $[N]$. Thus,

$$\begin{aligned}
\mathbb{E}_{\Omega|\mathcal{D}}[\phi(\Omega)] &= \mathbb{E}_{\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\Big(\mathcal{E}(\beta) - \widehat{\mathcal{E}}_{\Omega}(\beta)\Big)\right] \\
&= \mathbb{E}_{\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\Big(\mathbb{E}_{\Omega'}\Big[\widehat{\mathcal{E}}_{\Omega'}(\beta) - \widehat{\mathcal{E}}_{\Omega}(\beta)\Big]\Big)\right] \\
&\le \mathbb{E}_{\Omega,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\Big(\widehat{\mathcal{E}}_{\Omega'}(\beta) - \widehat{\mathcal{E}}_{\Omega}(\beta)\Big)\right] \\
&= \mathbb{E}_{\Omega,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\Big(\ell(\boldsymbol{A}_{i'_k}\beta^*; \widehat{\boldsymbol{A}}_{i'_k}\beta) - \ell(\boldsymbol{A}_{i_k}\beta^*; \widehat{\boldsymbol{A}}_{i_k}\beta)\Big)\right],
\end{aligned}$$

where the inequality follows by the convexity of the supremum function and Jensen's Inequality.

To proceed, we will use the ghost sampling technique. Recall that the entries of Ω and Ω' were drawn uniformly at random from $[N]$. As a result, $\ell(\boldsymbol{A}_{i'_k}\beta^*; \widehat{\boldsymbol{A}}_{i'_k}\beta) - \ell(\boldsymbol{A}_{i_k}\beta^*; \widehat{\boldsymbol{A}}_{i_k}\beta)$ and $\ell(\boldsymbol{A}_{i_k}\beta^*; \widehat{\boldsymbol{A}}_{i_k}\beta) -$

$\ell(\boldsymbol{A}_{i'_k}\beta^*;\widehat{\boldsymbol{A}}_{i'_k}\beta)$ have the same distribution. Further, since σ_k takes value 1 and -1 with equal probability, we have

$$\begin{aligned}&\mathbb{E}_{\Omega,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\Big(\ell(\boldsymbol{A}_{i'_k}\beta^*;\widehat{\boldsymbol{A}}_{i'_k}\beta)-\ell(\boldsymbol{A}_{i_k}\beta^*;\widehat{\boldsymbol{A}}_{i_k}\beta)\Big)\right]\\&\quad=\mathbb{E}_{\sigma,\Omega,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\sigma_k\Big(\ell(\boldsymbol{A}_{i'_k}\beta^*;\widehat{\boldsymbol{A}}_{i'_k}\beta)-\ell(\boldsymbol{A}_{i_k}\beta^*;\widehat{\boldsymbol{A}}_{i_k}\beta))\Big)\right].\end{aligned}$$

Combining the above relation with the fact that the supremum of a sum is bounded above by the sum of supremums, we obtain

$$\begin{aligned}\mathbb{E}_{\Omega|\mathcal{D}}[\phi(\Omega)]&\leq\mathbb{E}_{\sigma,\Omega,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\sigma_k\Big(\ell(\boldsymbol{A}_{i'_k}\beta^*;\widehat{\boldsymbol{A}}_{i'_k}\beta)-\ell(\boldsymbol{A}_{i_k}\beta^*;\widehat{\boldsymbol{A}}_{i_k}\beta)\Big)\right]\\&\leq\mathbb{E}_{\sigma,\Omega,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\sigma_k\ell(\boldsymbol{A}_{i'_k}\beta^*;\widehat{\boldsymbol{A}}_{i'_k}\beta)+\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}-\sigma_k\ell(\boldsymbol{A}_{i_k}\beta^*;\widehat{\boldsymbol{A}}_{i_k}\beta)\right]\\&=\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\sigma_k\ell(\boldsymbol{A}_{i_k}\beta^*;\widehat{\boldsymbol{A}}_{i_k}\beta)\right]+\mathbb{E}_{\sigma,\Omega'|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\frac{1}{n}\sum_{k=1}^{n}\sigma_k\ell(\boldsymbol{A}_{i'_k}\beta^*;\widehat{\boldsymbol{A}}_{i'_k}\beta)\right]\\&=2\cdot R_n(\ell\circ\mathcal{F}),\end{aligned}$$

where the second to last equality holds because σ_k is a symmetric random variable.

□

M.2.2 Lemma M.5

To prove Lemma M.5, we first prove a series of helper lemmas.

Lemma M.2. *Let Property 3.1 hold. Then, for any $\beta\in\mathcal{F}$,*

$$\max_{i\in[N]}\ell(\boldsymbol{A}_{i,\cdot}\beta^*,\widehat{\boldsymbol{A}}_{i,\cdot}\beta)\leq C(\widehat{\boldsymbol{A}}).$$

Here, $C(\widehat{\boldsymbol{A}})=2\Big[(B\cdot\|\widehat{\boldsymbol{A}}\|_\infty)^2+(\|\beta^*\|_1)^2\Big]$.

Proof. Observe that for any $i\in[N]$ and $\beta\in\mathcal{F}$,

$$\ell(\boldsymbol{A}_{i,\cdot}\beta^*,\widehat{\boldsymbol{A}}_{i,\cdot}\beta)=(\widehat{\boldsymbol{A}}_{i,\cdot}\beta-\boldsymbol{A}_{i,\cdot}\beta^*)^2\leq2(\widehat{\boldsymbol{A}}_{i,\cdot}\beta)^2+2(\boldsymbol{A}_{i,\cdot}\beta^*)^2.$$

Recall that every candidate vector $\beta \in \mathcal{F}$ has the following propery: $\|\beta\|_1 \leq B$. Hence, it follows that for any $i \in [N]$,

$$|\widehat{\boldsymbol{A}}_{i,\cdot}\widehat{\beta}| \leq \|\beta\|_1 \cdot \max_{j\in[p]} |\widehat{A}_{ij}| \ \leq B \cdot \|\widehat{\boldsymbol{A}}\|_\infty.$$

Further, By Property 3.1 and Holder's inequality, we have for any $i \in [N]$,

$$|\boldsymbol{A}_{i,\cdot}\beta^*| \leq \|\boldsymbol{A}_{i,\cdot}\|_\infty \|\beta^*\|_1 \leq \|\beta^*\|_1.$$

The desired result then follows from an immediate application of the above results. □

Lemma M.3. *Recall* $\operatorname{rank}(\widehat{\boldsymbol{A}}) = k$. *Then,*

$$R_n(\mathcal{F}) \leq \frac{\sqrt{k}B}{\sqrt{n}} \cdot \|\widehat{\boldsymbol{A}}\|_\infty.$$

Proof. Let $I_\beta = \{i \in [p] : \beta_i \neq 0\}$ denote the index set for the nonzero elements of $\beta \in \mathcal{F}$; recall that $|I_\beta| \leq k$ by the definition of $\mathcal{F}$. For any vector $v \in \mathbb{R}^p$, we denote v_{I_β} as the vector that retains only its values in I_β and takes the value 0 otherwise. Then,

$$\begin{aligned}
R_n(\mathcal{F}) &= \mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i \widehat{\boldsymbol{A}}_{i,\cdot}\beta\right)\right] \\
&= \frac{1}{n}\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\left(\sum_{j\in I_\beta}\beta_j\Big(\sum_{i=1}^n \sigma_i \widehat{\boldsymbol{A}}_{i,\cdot}\Big)_j\right)\right] \\
&\overset{(a)}{\leq} \frac{1}{n}\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\sup_{\beta\in\mathcal{F}}\|\beta\|_2 \cdot \left\|\Big(\sum_{i=1}^n \sigma_i \widehat{\boldsymbol{A}}_{i,\cdot}\Big)_{I_\beta}\right\|_2\right] \\
&\overset{(b)}{\leq} \frac{B}{n}\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\left\|\Big(\sum_{i=1}^n \sigma_i \widehat{\boldsymbol{A}}_{i,\cdot}\Big)_{I_\beta}\right\|_2\right] \\
&\overset{(c)}{\leq} \frac{B}{n}\left(\mathbb{E}_{\sigma,\Omega|\mathcal{D}}\left[\Big(\sum_{i=1}^n \sigma_i \widehat{\boldsymbol{A}}_{i,\cdot}\Big)_{I_\beta}\Big(\sum_{k=1}^n \sigma_k \widehat{\boldsymbol{A}}_{k,\cdot}\Big)_{I_\beta}^T\right]\right)^{1/2} \\
&= \frac{B}{n}\left(\mathbb{E}_{\Omega|\mathcal{D}}\left[\sum_{i=1}^n \left\|(\widehat{\boldsymbol{A}}_{i,\cdot})_{I_\beta}\right\|_2^2\right]\right)^{1/2} \\
&\leq \frac{B}{n}\left(nk\max_{i\in[n]}\left\|(\widehat{\boldsymbol{A}}_{i,\cdot})_{I_\beta}\right\|_\infty^2\right)^{1/2}
\end{aligned}$$

$$=\frac{\sqrt{k}B}{\sqrt{n}}\cdot\|\widehat{\boldsymbol{A}}\|_\infty.$$

Note that (a) makes use of the Cauchy-Schwartz Inequality, (b) follows from the boundedness assumption of the elements in $\mathcal{F}$ and noting the ℓ_2-norm of a vector is less than the ℓ_1-norm, and (c) applies Jensen's Inequality. □

Lemma M.4. Lipschitz composition of Rademacher averages. ([32])

Suppose $\{\phi_i\},\{\psi_i\}$, $i=1,\dots,n$, are two sets of functions on Θ such that for each i and $\theta,\theta'\in\Theta$, $|\phi_i(\theta)-\phi_i(\theta')|\le|\psi_i(\theta)-\psi_i(\theta')|$. Then, for all functions $c:\Theta\to\mathbb{R}$,

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left\{c(\theta)+\sum_{i=1}^n\sigma_i\phi_i(\theta)\right\}\right]\le\mathbb{E}\left[\sup_{\theta\in\Theta}\left\{c(\theta)+\sum_{i=1}^n\sigma_i\psi_i(\theta)\right\}\right],$$

where σ_i are Rademacher random variables.

Proof. The proof can be found in [32]. □

Lemma M.5. *Let Property 3.1 hold and recall* $\operatorname{rank}(\widehat{\boldsymbol{A}})=k$. *Then,*

$$R_n(\ell\circ\mathcal{F})\le C\frac{\sqrt{k}B^2}{\sqrt{n}}\cdot\|\widehat{\boldsymbol{A}}\|_\infty^2\cdot\|\beta^*\|_1,$$

where $C>0$ is an absolute constant.

Proof. Using Lemma M.2, we have for any $\beta\in\mathcal{F}$,

$$\max_{i\in[N]}|\ell'(\boldsymbol{A}_{i,\cdot}\beta^*,\widehat{\boldsymbol{A}}_{i,\cdot}\beta)|\le2\sqrt{C(\widehat{\boldsymbol{A}})},$$

where $\ell'(\cdot,\cdot)$ denotes the derivative of the loss function with respect to our estimate. Since our loss function of interest has bounded first derivative, the Lipschitz constant of $\ell(\cdot,\cdot)$ is bounded by $2C(\widehat{\boldsymbol{A}})^{1/2}$; hence, applying Lemma M.4 for Lipschitz functions and using Lemma M.3 yields the following inequality:

$$R_n(\ell\circ\mathcal{F})\le2\sqrt{C(\widehat{\boldsymbol{A}})}\cdot R_n(\mathcal{F})\le C\frac{\sqrt{k}B^2}{\sqrt{n}}\cdot\|\widehat{\boldsymbol{A}}\|_\infty^2\cdot\|\beta^*\|_1,$$

for some absolute constant, $C>0$. This concludes the proof. □

M.2.3 Proof of Theorem 3.2

Now we are ready to complete the proof of Theorem 3.2.

Proof of Theorem 3.2. The testing error, for PCR with parameter k or, equivalently, Linear Regression with covariate pre-processing via HSVT thresholded at the k-th singular value, is

$$\mathrm{MSE}(\widehat{Y}) = \frac{1}{N}\mathbb{E}_{\mathcal{D}|\Omega}\left[\sum_{i=1}^{N}\left(\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*\right)^2\right] = \mathbb{E}_{\mathcal{D}|\Omega}\left[\mathcal{E}(\widehat{\beta})\right],$$

where the expectation is taken with respect to the randomness in the data.

And, for a given training set Ω, the training error is

$$\mathrm{MSE}_{\Omega}(\widehat{Y}) = \frac{1}{n}\mathbb{E}_{\mathcal{D}|\Omega}\left[\sum_{i\in\Omega}\left(\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*\right)^2\right] = \mathbb{E}_{\mathcal{D}|\Omega}\left[\widehat{\mathcal{E}}_{\Omega}(\widehat{\beta})\right].$$

Recall that we shall consider the training set Ω being chosen uniformly at random amongst subsets of $[N]$ of size n. Given any Ω, observe that

$$\mathcal{E}(\widehat{\beta}) \le \widehat{\mathcal{E}}_{\Omega}(\widehat{\beta}) + \sup_{\beta\in\mathcal{F}}\left(\mathcal{E}(\beta) - \widehat{\mathcal{E}}_{\Omega}(\beta)\right) = \widehat{\mathcal{E}}_{\Omega}(\widehat{\beta}) + \phi(\Omega),$$

where $\phi(\Omega)$ is as defined by (63). Taking expectations of the above inequality, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{D},\Omega}[\mathcal{E}(\widehat{\beta})] &\le \mathbb{E}_{\mathcal{D},\Omega}[\widehat{\mathcal{E}}(\widehat{\beta})] + \mathbb{E}_{\mathcal{D},\Omega}[\phi(\Omega)] \\ &= \mathbb{E}_{\Omega}\left[\mathbb{E}_{\mathcal{D}|\Omega}[\widehat{\mathcal{E}}(\widehat{\beta})]\right] + \mathbb{E}_{\mathcal{D}}\left[\ \mathbb{E}_{\Omega|\mathcal{D}}[\phi(\Omega)]\right]. \qquad (64)\end{aligned}$$

We now bound each term on the right-hand side of (64) separately. Beginning with the leftmost term, observe that, by definition, we have

$$\mathbb{E}_{\Omega}\left[\mathbb{E}_{\mathcal{D}|\Omega}[\widehat{\mathcal{E}}(\widehat{\beta})]\right] = \mathbb{E}_{\Omega}\left[\mathrm{MSE}_{\Omega}(\widehat{Y})\right].$$

Moreover, applying Lemmas M.1 and M.5, we obtain

$$\mathbb{E}_{\mathcal{D}}\left[\ \mathbb{E}_{\Omega|\mathcal{D}}[\phi(\Omega)]\right] \le 2\mathbb{E}_{\mathcal{D}}[R_n(\ell\circ\mathcal{F})]$$

$$\leq C\mathbb{E}_{\mathcal{D}}\left[\frac{\sqrt{k}B^2}{\sqrt{n}}\cdot\|\widehat{\boldsymbol{A}}\|_\infty^2\cdot\|\beta^*\|_1\right]$$
$$=CC''\frac{k^{5/2}}{\sqrt{n}}\mathbb{E}_{\mathcal{D}}\left[\|\widehat{\beta}\|^2\cdot\|\widehat{\boldsymbol{A}}\|_\infty^2\cdot\right]\cdot\|\beta^*\|_1$$

where we recall $B:=C''\cdot k\cdot\|\widehat{\beta}\|_1$. Combining the above results completes the proof. □

N Proof of Propositions 3.1, 3.2 and N.1: Examples

N.1 Proof of Proposition 3.1: Embedded Random Gaussian Features

Recall that we let $\boldsymbol{A}=\tilde{\boldsymbol{A}}\tilde{\boldsymbol{R}}$ where $\tilde{\boldsymbol{A}}\in\mathbb{R}^{N\times r}$ is a random matrix whose entries are independent standard normal random variables, i.e., $\tilde{A}_{ij}\sim\mathcal{N}(0,1)$, and $\tilde{\boldsymbol{R}}\in\mathbb{R}^{r\times p}$ is another random matrix with independent entries such that $\tilde{R}_{ij}=1/\sqrt{r}$ with probability 1/2 and $\tilde{R}_{ij}=-1/\sqrt{r}$ with probability 1/2 in Proposition 3.1. In this subsection, we show that $s_r(\boldsymbol{A})=\Omega\left(\sqrt{\frac{Np}{r}}\right)$ and $\|\boldsymbol{A}\|_\infty=O(\sqrt{\log(Np)})$ with high probability.

N.1.1 Helper Lemmas

Lemma N.1. *Suppose that $r\leq\frac{\sqrt{p}}{4\sqrt{2\log p}}+1$ and let $\boldsymbol{R}\in\mathbb{R}^{r\times p}$ be a random matrix with independent entries such that $\boldsymbol{R}_{ij}=\frac{1}{\sqrt{p}}$ with probability $\frac{1}{2}$ and $\boldsymbol{R}_{ij}=-\frac{1}{\sqrt{p}}$ with probability $\frac{1}{2}$. With probability at least $1-\frac{1}{p^2}$, for all $v\in\mathbb{R}^r$,*

$$\frac{1}{2}\|v\|_2^2\leq\|\boldsymbol{R}^Tv\|_2^2\leq\frac{3}{2}\|v\|_2^2.$$

Proof. For $i\in[r]$, let $\boldsymbol{R}_i$ denote the i-th row of $\boldsymbol{R}$. Observe that $\|\boldsymbol{R}_i\|_2=1$ for all $i\in[r]$. Also, note that for $i\neq j\in[r]$, $\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle=\frac{1}{p}\sum_{k=1}^p\tilde{\boldsymbol{R}}_{ik}\tilde{\boldsymbol{R}}_{jk}$ is a sum of p independent binary random variables; $\tilde{\boldsymbol{R}}_{ik}\tilde{\boldsymbol{R}}_{jk}=1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. Therefore, $\mathbb{E}\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle=0$. By Hoeffding's inequality for bounded random variables,

$$\mathbb{P}(|\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle|>t)\leq2\exp\left(-\frac{pt^2}{2}\right).$$

Letting $t=\frac{2\sqrt{2\log p}}{\sqrt{p}}$, we can conclude that for any pair of $i\neq j\in[r]$, $|\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle|\leq\frac{2\sqrt{2\log p}}{\sqrt{p}}$ with probability at least $1-\frac{2}{p^4}$. There are $\binom{r}{2}\leq\frac{r^2}{2}$ such pairs and $r\leq p$. Thus, applying the union bound, we know that $|\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle|\leq\frac{2\sqrt{2\log p}}{\sqrt{p}}$ for all pairs $i\neq j$ with probability at least $1-\frac{1}{p^2}$.

Now we observe that

$$\begin{aligned}\|\boldsymbol{R}^Tv\|_2^2&=\left\langle\sum_{i=1}^r v_i\boldsymbol{R}_i,\sum_{i=1}^r v_i\boldsymbol{R}_i,\right\rangle\\&=\sum_{i=1}^r v_i^2\|\boldsymbol{R}_i\|_2^2+\sum_{i=1}^r\sum_{j\neq i}v_iv_j\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle\\&\leq\sum_{i=1}^r v_i^2\|\boldsymbol{R}_i\|_2^2+\sum_{i=1}^r\sum_{j\neq i}|v_iv_j||\langle\boldsymbol{R}_i,\boldsymbol{R}_j\rangle|.\end{aligned}$$

With probability at least $1-\frac{1}{p^2}$,

$$\begin{aligned}\|\boldsymbol{R}^Tv\|_2^2&\leq\sum_{i=1}^r v_i^2\|\boldsymbol{R}_i\|_2^2+\sum_{i=1}^r\sum_{j\neq i}|v_iv_j|\frac{2\sqrt{2\log p}}{\sqrt{p}}\\&\overset{(a)}{\leq}\sum_{i=1}^r v_i^2+(r-1)\sum_{i=1}^r v_i^2\frac{2\sqrt{2\log p}}{\sqrt{p}}\\&\leq\|v\|_2^2\left(1+\frac{2(r-1)\sqrt{2\log p}}{\sqrt{p}}\right)\end{aligned}$$

where (a) follows from that $\|\boldsymbol{R}_i\|_2^2=1$ for all $i\in[r]$ and the Cauchy-Schwarz inequality $(2|v_iv_j|\leq v_i^2+v_j^2)$. By the same argument, $\|\boldsymbol{R}^Tv\|_2^2\geq\|v\|_2^2\left(1-\frac{2(r-1)\sqrt{2\log p}}{\sqrt{p}}\right)$.

Lastly, we note that $\frac{2(r-1)\sqrt{2\log p}}{\sqrt{p}}\leq\frac{1}{2}$ if and only if $r\leq\frac{\sqrt{p}}{4\sqrt{2\log p}}+1$ to complete the proof. □

Remark N.1. *Lemma N.1 implies that given $r\leq1+\frac{\sqrt{p}}{4\sqrt{2\log p}}$, the right multiplication of $\boldsymbol{R}$ defines a quasi-isometric embedding from $\mathbb{R}^r$ to $\mathbb{R}^p$ with high probability. More precisely, with probability at least $1-\frac{1}{p^2}$, the following inequalities are true:*

$$\frac{1}{2}\|v\|_2^2\leq\|\boldsymbol{R}^Tv\|_2^2\leq\frac{3}{2}\|v\|_2^2,\quad\forall v\in\mathbb{R}^r,\qquad\textit{and}\qquad\frac{1}{2}\|w\|_2^2\leq\|\boldsymbol{R}w\|_2^2\leq\frac{3}{2}\|w\|_2^2,\quad\forall w\in\textit{rowspace}(\boldsymbol{R}).$$

The first inequality is just the conclusion of Lemma N.1; it implies that $\frac{1}{2}\leq\lambda_i(\boldsymbol{R}\boldsymbol{R}^T)\leq\frac{3}{2}$ for all $i\in[r]$ where $\lambda_i(\boldsymbol{R}\boldsymbol{R}^T)$ denotes the i-th largest eigenvalue of $\boldsymbol{R}\boldsymbol{R}^T$. Let v_i be an eigenvector corresponding to $\lambda_i(\boldsymbol{R}\boldsymbol{R}^T)$; $\{v_1,\ldots,v_r\}$ forms an orthonormal basis of $\mathbb{R}^r$.

To see why the second inequality also holds, suppose that $w = \boldsymbol{R}^T v_w$ *for some* $v_w \in \mathbb{R}^r$ *(such a* v_w *exists because* $w \in \boldsymbol{R}^T$*). Observe that* $\|w\|_2^2 = w^T w = v_w^T \boldsymbol{R}\boldsymbol{R}^T v_w$ *and that* $\|\boldsymbol{R}w\|_2^2 = w^T \boldsymbol{R}^T \boldsymbol{R} w = v_w^T \boldsymbol{R}\boldsymbol{R}^T \boldsymbol{R}\boldsymbol{R}^T v_w$*. We may write* $v_w = \sum_{i=1}^r c_i v_i$ *for some* $c_i \in \mathbb{R}$*. It follows that* $\|w\|_2^2 = \sum_{i=1}^r c_i^2 \lambda_i(\boldsymbol{R}\boldsymbol{R}^T)$ *and* $\|\boldsymbol{R}w\|_2^2 = \sum_{i=1}^r c_i^2 \lambda_i^2(\boldsymbol{R}\boldsymbol{R}^T)$*; therefore,* $\frac{1}{2} \leq \lambda_r(\boldsymbol{R}\boldsymbol{R}^T) \leq \frac{\|\boldsymbol{R}w\|_2^2}{\|w\|_2^2} \leq \lambda_1(\boldsymbol{R}\boldsymbol{R}^T) \leq \frac{3}{2}$.

Remark N.2. *By Remark N.1, with probability at least* $1 - \frac{1}{p^2}$,

$$\begin{aligned} s_r(\tilde{\boldsymbol{A}}\boldsymbol{R}) &= \sup_{\substack{W \subset \mathbb{R}^p \\ \dim W = r}} \inf_{w \in W} \frac{\|\tilde{\boldsymbol{A}}\boldsymbol{R}w\|_2}{\|w\|_2} = \inf_{w \in \text{rowspace}\boldsymbol{R}} \frac{\|\tilde{\boldsymbol{A}}\boldsymbol{R}w\|_2}{\|w\|_2} \\ &\geq \sqrt{\frac{1}{2}} \inf_{w \in \text{rowspace}\boldsymbol{R}} \frac{\|\tilde{\boldsymbol{A}}\boldsymbol{R}w\|_2}{\|\boldsymbol{R}w\|_2} = \sqrt{\frac{1}{2}} \inf_{v \in \mathbb{R}^r} \frac{\|\tilde{\boldsymbol{A}}v\|_2}{\|v\|_2} = \sqrt{\frac{1}{2}} s_r(\tilde{\boldsymbol{A}}). \end{aligned}$$

Lemma N.2 (Spectral properties of $\tilde{\boldsymbol{A}}$). *Let* $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times r}$ *be a random matrix whose entries are i.i.d. standard Gaussian random variable. Then,*

(1) with probability at least $1 - 2\exp\left(-\frac{1}{2}\sqrt{Nr}\right)$, $\text{rank}(\tilde{\boldsymbol{A}}) = r$ *and*

$$\frac{s_1(\tilde{\boldsymbol{A}})}{s_r(\tilde{\boldsymbol{A}})} \leq \frac{1 + (r/N)^{1/4} + (r/N)^{1/2}}{1 - (r/N)^{1/4} - (r/N)^{1/2}};$$

(2) with probability at least $1 - \exp\left(-\frac{Nr}{8}\right)$,

$$\|\tilde{\boldsymbol{A}}\|_F^2 > \frac{Nr}{2}.$$

Proof. **Proof of Claim 1** By [44, Corollary 5.35], for any $t \geq 0$, we have

$$\sqrt{N} - \sqrt{r} - t \leq s_{\min}(\tilde{\boldsymbol{A}}) \leq s_{\max}(\tilde{\boldsymbol{A}}) \leq \sqrt{N} + \sqrt{r} + t,$$

with probability at least $1 - 2\exp(-t^2/2)$. Choosing $t = (Nr)^{1/4}$ concludes the proof.

Proof of Claim 2 Observe that $\|\tilde{\boldsymbol{A}}\|_F^2 = \sum_{i,j} \tilde{\boldsymbol{A}}_{ij}^2$. We can easily observe that $\mathbb{E}\|\tilde{\boldsymbol{A}}\|_F^2 = Nr$. By Bernstein's inequality, it follows that for every $t \geq 0$,

$$\mathbb{P}\{\|\tilde{\boldsymbol{A}}\|_F^2 - \mathbb{E}\|\tilde{\boldsymbol{A}}\|_F^2 \leq -t\} \leq \exp\left(-\frac{1}{2}\min\left\{\frac{t^2}{Nr}, t\right\}\right).$$

With $t=\frac{Nr}{2}$, we have

$$\mathbb{P}\{\|\tilde{\boldsymbol{A}}\|_F^2\leq\frac{Nr}{2}\}\leq\exp\left(-\frac{Nr}{8}\right).$$

□

Remark N.3. *Lemma N.2 implies that with probability at least* $1-2\exp\left(-2\sqrt{Nr}\right)-\exp\left(-\frac{Nr}{8}\right)$,

$$s_r(\tilde{\boldsymbol{A}})^2\geq\left[1+(r-1)\frac{s_1(\tilde{\boldsymbol{A}})^2}{s_r(\tilde{\boldsymbol{A}})^2}\right]^{-1}\|\tilde{\boldsymbol{A}}\|_F^2\geq\left[1+(r-1)\left(\frac{1+(r/N)^{1/4}+(r/N)^{1/2}}{1-(r/N)^{1/4}-(r/N)^{1/2}}\right)^2\right]^{-1}\frac{Nr}{2}.$$

Lemma N.3 (Structural properties of $\boldsymbol{A}$). *Let $\boldsymbol{A}\in\mathbb{R}^{N\times p}$ be a matrix generated as above. With probability at least $1-\frac{2}{N^2p}$,*

$$\max_{i,j}|A_{ij}|\leq 4\sqrt{\log(Np)}.$$

Proof. By construction, $A_{ij}=\sum_{k=1}^{r}\tilde{A}_{ik}\tilde{R}_{kj}$ and $A_{ij}|\tilde{\boldsymbol{R}}\sim\mathcal{N}(0,\sum_{k=1}^{r}\tilde{R}_{kj}^2)$ conditioned on $\tilde{\boldsymbol{R}}$ and note $\sum_{k=1}^{r}\tilde{R}_{kj}^2=1$ regardless of $\tilde{\boldsymbol{R}}$. Therefore, for each fixed $j\in[p]$, $A_{\cdot j}|\tilde{\boldsymbol{R}}\sim\mathcal{N}(0,I_N)$. Observe that $\max_i|A_{ij}|\Big|\tilde{\boldsymbol{R}}$ is the maximum absolute value of N i.i.d. standard Gaussians and $\mathbb{E}\Big[\max_i|A_{ij}|\Big|\tilde{\boldsymbol{R}}\Big]\leq 2\sqrt{\log N}$. Since this holds regardless of $\tilde{\boldsymbol{R}}$, by tower law we can remove the conditioning on $\tilde{\boldsymbol{R}}$. In addition, by the concentration of Lipschitz function (note that $\max:\mathbb{R}^N\rightarrow\mathbb{R}$ is 1-Lipschitz),

$$\mathbb{P}\left(|\max_i|A_{ij}|-\mathbb{E}[\max_i|A_{ij}|]|\geq t\right)\leq 2\exp\left(-\frac{t^2}{2}\right).$$

Letting $t=2\sqrt{\log(Np)}$, it follows for each $j\in[p]$ that $\mathbb{P}\left(|\max_i|A_{ij}|\geq 4\sqrt{\log(Np)}\right)\leq\frac{2}{N^2p^2}$. Taking union bound over $j\in[p]$, we conclude that with probability at least $1-\frac{2}{N^2p}$,

$$\max_{i,j}|A_{ij}|\leq 4\sqrt{\log(Np)}.$$

□

N.1.2 Completing the Proof of Proposition 3.1

Proof of Proposition 3.1. Observe that $\boldsymbol{A} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{R}} = \sqrt{\frac{p}{r}}\tilde{\boldsymbol{A}}\boldsymbol{R}$. By Lemmas N.1, N.2 (along with Remarks N.2 and N.3), we have

$$s_r(\boldsymbol{A}) = s_r(\tilde{\boldsymbol{A}}\tilde{\boldsymbol{R}}) = \sqrt{\frac{p}{r}} s_r(\tilde{\boldsymbol{A}}\boldsymbol{R}) \geq \sqrt{\frac{Np}{4r}} \left[\Delta + \frac{1}{r}(1-\Delta)\right]^{-1/2}$$

with probability at least $1 - 2\exp\left(-2\sqrt{Nr}\right) - \exp\left(-\frac{Nr}{8}\right)$, where $\Delta = \frac{1+(r/N)^{1/4}+(r/N)^{1/2}}{1-(r/N)^{1/4}-(r/N)^{1/2}}$. Note that if $r \ll N$, then $|\Delta - 1| = o(1)$. This inequality combined with Lemma N.3 completes the proof.

□

N.2 Proof of Proposition 3.2: Geometrically Decaying Singular Values

Proof of Proposition 3.2. Recall the (slightly simplified) bound of Corollary 3.2 is

$$MSE_{\Omega}(\widehat{Y}) \leq \frac{C'\|\beta^*\|_1^2}{\rho^4}\left(\frac{k}{n} + \frac{n \vee p}{(\tau_k - \tau_{k+1})^2}\right)\log^5(np) + \frac{3\|\beta^*\|_1^2}{n}\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2 + \frac{20}{n}\|\phi\|_2^2, \tag{65}$$

where $C' = C(1+\sigma^2)(1+\gamma^2)(1+K_\alpha^4)$ and $C > 0$ is an absolute constant.

Let us evaluate each of the first four terms in the right hand side of (65) to reach the desired (7).

First term. Due to choice of k we immediately have follows that it is

$$\frac{C'\|\beta^*\|_1^2}{\rho^4}\log^5(np)\frac{k}{n} \leq \frac{C'C'(\theta)\|\beta^*\|_1^2}{\rho^4}\frac{C_2\log^6(np)}{n}.$$

Second term.

$$\begin{aligned}
\frac{C'\|\beta^*\|_1^2}{\rho^4}\frac{n \vee p}{(\tau_k - \tau_{k+1})^2}\log^5(np) &\leq \frac{C'\|\beta^*\|_1^2}{\rho^4}\frac{n \vee p}{(\sqrt{Np}(\theta^{k-1} - \theta^k))^2}\log^5(np) \\
&= \frac{C'\|\beta^*\|_1^2}{\rho^4}\frac{n \vee p}{Np(\theta^{k-1}(1-\theta))^2}\log^5(np) \\
&\leq \frac{C'\|\beta^*\|_1^2}{\rho^4}C(\theta)\frac{1}{n \wedge p}\frac{1}{\theta^{2k}}\log^5(np)
\end{aligned}$$

$$\leq \frac{C'\|\beta^*\|_1^2}{\rho^4} C(\theta) \frac{1}{(n \wedge p)^{1/2}} \log^5(np)$$

where we have used the fact that $\tau_i = \tau_1 \theta^{i-1}$ for $i \geq 1$, $\tau_1 = C_1\sqrt{Np}$, $n = \Theta(N)$ and $C(\theta) > 0$ is a term that depends only on θ.

Third term. The goal is to bound $\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2$. With notation $\boldsymbol{E} = \boldsymbol{A} - \boldsymbol{A}^k$, this is equivalent to bounding $\max_{j \in [p]} \|\boldsymbol{E}_{\cdot,j}\|_2^2$. With $\boldsymbol{A} = \sum_{i=1}^N \tau_i \mu_i \nu_i^T$ where $\mu_i \in \mathbb{R}^N$, $\nu_i \in \mathbb{R}^p$ for $i \in [N]$, for any $j \in [p]$, we have

$$\begin{aligned}
\frac{1}{n}\|\boldsymbol{E}_{\cdot,j}\|^2 = \frac{1}{n}\left\|\left(\sum_{i=k+1}^N \tau_i \mu_i \nu_i^T\right) e_j\right\|^2 &= \frac{1}{n}\left\|\sum_{i=k+1}^N \tau_i \mu_i (\nu_i^T e_j)\right\|^2 \\
&\overset{(a)}{=} \frac{1}{n} \sum_{i=k+1}^N \tau_i^2 (\nu_i^T e_j)^2 \\
&\overset{(b)}{\leq} \frac{1}{n} \sum_{i=k+1}^N \tau_1^2 \theta^{2(i-1)} (\nu_i^T e_j)^2 \\
&\overset{(c)}{\leq} \frac{C_1 N p}{n} \sum_{i=k+1}^N \theta^{2(i-1)} (\nu_i^T e_j)^2 \\
&\overset{(d)}{\leq} \frac{C_1 N p}{np} \sum_{i=k+1}^N \theta^{2(i-1)} \\
&\overset{(e)}{\leq} C\theta^{2k} \overset{(f)}{\leq} \frac{C}{(n \wedge p)^{1/2}}
\end{aligned}$$

Here, (a) follows from the orthonormality of the (left) singular vectors; (b) follows from $\tau_i = \tau_1 \theta^{i-1}$; (c) follows from $\tau_1 = C_1\sqrt{Np}$; (d) 'incoherence' property of singular vector, i.e. $\nu_i^T e_j = O(1/\sqrt{p})$ for all $i,j \in [p]$; (e) follows from property of geometric series for some absolute constant $C > 0$; and (f) follows from choice of k.

Concluding the proof. The final term is repeat of $\frac{20}{n}\|\phi\|_2^2$. Therefore, putting all of the above together, the proof concludes. □

N.3 Geometrically Decaying Singular Values - Example from Signal Processing

As an illustration, we construct a matrix, popular in signal processing, which satisfies the conditions on the spectrum laid out in Proposition 7. We will construct an example based on the incoherence between the canonical basis and the Discrete Fourier Transform (DFT) basis.

Suppose that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where: (i) $\boldsymbol{\Sigma}$ is a diagonal matrix such that $\Sigma_{11} = C\sqrt{Np}$ for some $C > 0$ and the diagonal entries of $\boldsymbol{\Sigma}$ satisfy $0 \leq \Sigma_{i+1,i+1}/\Sigma_{i,i} \leq \theta$ for all $i \in [N \wedge p - 1]$ and for some $\theta \in (0,1)$; (ii) $\boldsymbol{U} \in \mathbb{R}^{N \times N}$ is a DFT matrix such that $U_{ij} = (1/\sqrt{N}) \cdot \exp(2\pi \boldsymbol{i}(i-1)(j-1)/N)$ for all $i,j \in [N]$, where $\boldsymbol{i}$ denotes the imaginary unit; (iii) $\boldsymbol{V} \in \mathbb{R}^{p \times p}$ is a DFT matrix such that $V_{ij} = (1/\sqrt{p}) \cdot \exp(2\pi \boldsymbol{i}(i-1)(j-1)/p)$ for all $i,j \in [p]$.

The entries of the resulting matrix $\boldsymbol{A}$ are complex numbers, but one could also construct $\boldsymbol{A}$ by taking $\boldsymbol{U}$ and $\boldsymbol{V}$ as discrete cosine (or sine) transform matrices. Further, observe that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices; hence, $\sigma_i(\boldsymbol{A}) = \sigma_i(\boldsymbol{\Sigma})$ for all $i \in [N \wedge p]$. Finally, to show $\boldsymbol{A}$ fits within our setting, we argue $\|\boldsymbol{A}\|_\infty \leq C'$ for some constant $C' > 0$.

Proposition N.1. *Let $\boldsymbol{A}$ be generated as above. Then, $\|\boldsymbol{A}\|_\infty \leq C/(1-\theta)$. Here, $C > 0$ and $\theta \in (0,1)$ are the constants that appear in the description of $\boldsymbol{\Sigma}$. Further, we have $v_i^T e_j = O(1/\sqrt{p})$ for all $i,j \in [p]$.*

Proof of Proposition N.1 can be found in Appendix N.3.

Proof of Proposition N.1. For $(i,j) \in [N] \times [p]$, we have $\boldsymbol{A}_{ij} = \sum_{k=1}^{N \wedge p} \Sigma_{kk} U_{ik} V_{jk}$. Thus,

$$\begin{aligned} |A_{ij}| &= \left| \sum_{k=1}^{N \wedge p} \Sigma_{kk} U_{ik} V_{jk} \right| \leq \sum_{k=1}^{N \wedge p} \Sigma_{kk} |U_{ik}| |V_{jk}| \\ &\overset{(a)}{\leq} \sum_{k=1}^{N \wedge p} \Sigma_{11} \theta^{k-1} \frac{1}{\sqrt{Np}} = \Sigma_{11} \frac{1-\theta^{N \wedge p}}{1-\theta} \frac{1}{\sqrt{Np}} \\ &\overset{(b)}{\leq} \frac{C}{1-\theta}. \end{aligned}$$

Here, (a) follows from that $|U_{ik}|=\frac{1}{\sqrt{N}}$, $|V_{jk}|=\frac{1}{\sqrt{p}}$, and $\Sigma_{kk}\leq\Sigma_{11}\theta^{k-1}$; and (b) follows from the assumption $\Sigma_{11}=C\sqrt{Np}$ and that $1-\theta^{N\wedge p}\leq 1$. □

O Proof of Propositions 3.3, 4.1

O.1 Proof of Proposition 3.3

This analysis is taken from [48] and is stated for completeness.

Step 1: Partitioning the space $[0,1)^K$. Let $\mathcal{E}$ denote a partition of the cube $[0,1)^K$ into a finite number (denoted by $|\mathcal{E}|$) of cubes Δ. Let $\ell\in\mathbb{N}$. We say $P_{\mathcal{E},\ell}:[0,1]^K\rightarrow\mathbb{R}$ is a piecewise polynomial of degree ℓ if

$$P_{\mathcal{E},\ell}(\theta)=\sum_{\Delta\in\mathcal{E}}P_{\Delta,\ell}(\theta)\mathbb{1}(\theta\in\Delta), \tag{66}$$

where $P_{\Delta,\ell}(\theta):[0,1]^K\rightarrow\mathbb{R}$ denotes a polynomial of degree at most ℓ.

It suffices to consider an equal partition of $[0,1)^K$. More precisely, for any $k\in\mathbb{N}$, we partition the the set $[0,1)$into $1/k$ half-open intervals of lengths $1/k$, i.e, $[0,1)=\cup_{i=1}^{k}[(i-1)/k,i/k)$. It follows that $[0,1)^K$ can be partitioned into k^K cubes of forms $\otimes_{j=1}^{K}[(i_j-1)/k,i_j/k)$ with $i_j\in[k]$. Let $\mathcal{E}_k$ be such a partition with $I_1,I_2,...,I_{k^K}$ denoting all such cubes and $z_1,z_2,...,z_{k^K}\in\mathbb{R}^K$ denoting the centers of those cubes.

Step 2: Taylor Expansion of $g(\cdot,\rho_j)$. For Step 2 of the proof, to reduce notational overload, we suppress dependence of ρ_j on g, i.e.let $g(\cdot)=g(\cdot,\rho_j)$.

For every I_i with $1\leq i\leq k^K$, define $P_{I_i,\ell}(\theta)$ as the degree-ℓ Taylor's series expansion of $g(\theta)$ at point z_i:

$$P_{I_i,\ell}(\theta)=\sum_{\kappa:|\kappa|\leq\ell}\frac{1}{\kappa!}(\theta-z_i)^{\kappa}\nabla_{\kappa}g(z_i),$$

where $\kappa=(\kappa_1,...,\kappa_K)$ is a multi-index with $\kappa!=\prod_{i=1}^{d}\kappa_i!$, and $\nabla_{\kappa}g(z_i)$ is the partial derivative defined in (10). Note similar to g, $P_{I_i,\ell}(\theta)$ really refers to $P_{I_i,\ell}(x,\rho_j)$ Now we define a degree-ℓ piecewise polynomial as in (66), i.e.,

$$P_{\mathcal{E}_k,\ell}(\theta) = \sum_{i=1}^{k^K} P_{I_i,\ell}(\theta)\mathbb{1}(\theta \in I_i).$$

For the remainder of the proof, let $\ell = \lfloor \zeta \rfloor$. Since $g(\cdot,\rho_j) \in \mathcal{H}(\zeta,L)$, it follows from the that

$$\begin{aligned}
&\sup_{\theta\in[0,1)^K} |g(\theta) - P_{\mathcal{E}_k,\ell}(\theta)| \\
&= \sup_{1\le i\le k^K} \sup_{\theta\in I_i} |g(\theta) - P_{I_i,\ell}(\theta)| \\
&\overset{(a)}{=} \sup_{1\le i\le k^K} \sup_{\theta\in I_i} \left| \sum_{\kappa:|\kappa|\le \ell-1} \frac{\nabla_\kappa g(z_i)}{\kappa!}(\theta - z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i^{'})}{\kappa!}(\theta - z_i)^\kappa - P_{I_i,\ell}(\theta) \right| \\
&= \sup_{1\le i\le k^K} \sup_{\theta\in I_i} \left| \sum_{\kappa:|\kappa|\le \ell-1} \frac{\nabla_\kappa g(z_i)}{\kappa!}(\theta - z_i)^\kappa \pm \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i)}{\kappa!}(\theta - z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i^{'})}{\kappa!}(\theta - z_i)^\kappa - P_{I_i,\ell}(\theta) \right| \\
&= \sup_{1\le i\le k^K} \sup_{\theta\in I_i} \left| \sum_{\kappa:|\kappa|\le \ell} \frac{\nabla_\kappa g(z_i)}{\kappa!}(\theta - z_i)^\kappa + \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i^{'}) - \nabla_\kappa g(z_i)}{\kappa!}(\theta - z_i)^\kappa - P_{I_i,\ell}(\theta) \right| \\
&= \sup_{1\le i\le k^K} \sup_{\theta\in I_i} \left| \sum_{\kappa:|\kappa|=\ell} \frac{\nabla_\kappa g(z_i^{'}) - \nabla_\kappa g(z_i)}{\kappa!}(\theta - z_i)^\kappa \right| \\
&\overset{(b)}{\le} \sup_{1\le i\le k^K} \sup_{\theta\in I_i} \|\theta - z_i\|_\infty^\ell \sup_{\theta\in I_i} \sum_{\kappa:|\kappa|=\ell} \frac{1}{\kappa!} \left| \nabla_\kappa g(z_i^{'}) - \nabla_\kappa g(z_i) \right| \\
&\overset{(c)}{\le} \mathcal{L} \sup_{1\le i\le k^K} \sup_{\theta\in I_i} \|\theta - z_i\|_\infty^\zeta = \mathcal{L} k^{-\zeta}.
\end{aligned}$$

where (a) follows from multivariate's version of Taylor's theorem (and using the Lagrange form for the remainder) and $z_i^{'} \in [0,1)^K$ is a vector that can be represented as $z_i^{'} = (1-c)z_i + cx$ for $c \in (0,1)$; (b) follows from Holder's inequality; (c) follows from Definition 3.1.

Step 3: Construct Low-Rank Approximation of $\boldsymbol{A}'$ Using $P_{\mathcal{E}_k,\ell}(\cdot,\rho_j)$. Recall $\boldsymbol{A}'_{ij} = g(\theta_i,\rho_j)$, and $g(\cdot,\rho_j) \in \mathcal{H}(\zeta,\mathcal{L})$. We now construct a low-rank approximation of it using $P_{I_i,\ell}(\cdot,\rho_j)$. Define $\boldsymbol{A}^{(\mathrm{lr})} \in \mathbb{R}^{N\times p}$, where $\boldsymbol{A}^{(\mathrm{lr})}_{ij} = P_{\mathcal{E}_k,\ell}(\theta_i,\rho_j)$.

By Step 2, we have that for all $i \in [N], j \in [p]$,

$$\left| \boldsymbol{A}'_{ij} - \boldsymbol{A}^{(\mathrm{lr})}_{ij} \right| \le \mathcal{L} k^{-\zeta}$$

It remains to bound the rank of $\boldsymbol{A}^{(\mathrm{lr})}$. Note that since $P_{\mathcal{E}_k,\ell}(\theta_i,\rho_j)$ is a piecewise polynomial of degree $\ell = \lfloor \zeta \rfloor$, it has a decomposition of the form

$$\boldsymbol{A}^{(\mathrm{lr})}_{ij} = P_{\mathcal{E}_k,\ell}(\theta_i,\rho_j) = \sum_{i=1}^{k^K} \langle \Phi(\theta), \beta_{I_i,s} \rangle \mathbb{1}(\theta \in I_i)$$

where the vector

$$\Phi(\theta) = \left(1,\theta_1,\dots,\theta_K,\dots,\theta_1^\ell,\dots,\theta_K^\ell\right)^T,$$

i.e., is the vector of all monomials of degree less than or equal to ℓ. The number of such monomials is easily show to be equal to $C(\zeta,K) := \sum_{i=0}^{\lfloor \zeta \rfloor} \binom{i+K-1}{K-1}$.

Thus the rank of $\boldsymbol{A}^{(\mathrm{lr})}$ is bounded by $k^K C(\zeta,K)$. Setting $k = 1/\delta$ completes the proof.

O.2 Proof of Proposition 4.1

Let $\boldsymbol{A}^{(\mathrm{lr})}$ and β^* be defined as in Property 4.1. Then,

$$\begin{aligned}
|A'_{i0} - \sum_{k=1}^r \beta_k^* \cdot A'_{ik}| &= |A'_{i0} \pm \boldsymbol{A}^{(\mathrm{lr})}_{i0} - \sum_{k=1}^r \beta_k^* \cdot A'_{ik} \pm \sum_{k=1}^r \beta_k^* \cdot \boldsymbol{A}^{(\mathrm{lr})}_{ik}| \\
&\le |A'_{i0} - \boldsymbol{A}^{(\mathrm{lr})}_{i0}| + |\sum_{k=1}^r \beta_k^* \cdot A'_{ik} - \sum_{k=1}^r \beta_k^* \cdot \boldsymbol{A}^{(\mathrm{lr})}_{ik}| + |\boldsymbol{A}^{(\mathrm{lr})}_{i0} - \sum_{k=1}^r \beta_k^* \cdot \boldsymbol{A}^{(\mathrm{lr})}_{ik}| \\
&= |A'_{i0} - \boldsymbol{A}^{(\mathrm{lr})}_{i0}| + |\sum_{k=1}^r \beta_k^* \cdot A'_{ik} - \sum_{k=1}^r \beta_k^* \cdot \boldsymbol{A}^{(\mathrm{lr})}_{ik}| \\
&\le |A'_{i0} - \boldsymbol{A}^{(\mathrm{lr})}_{i0}| + \sum_{k=1}^r |\beta_k^* \cdot A'_{ik} - \beta_k^* \cdot \boldsymbol{A}^{(\mathrm{lr})}_{ik}| \\
&\le C(r+1)\boldsymbol{\mathcal{L}} \cdot \delta^\zeta
\end{aligned}$$

By Property 4.1, we have $r \le C(\zeta,K)\left(\frac{1}{\delta}\right)^K$, which completes the proof.

O.3 Proof of Theorem 4.1

Proof. The bound in Theorem 4.1, given by (19), is a sum of the pre-intervention error term and the additional penalty paid for the generalization error in the post-intervention period. The first term,

$$\frac{C'C(\zeta,K)\mathcal{L}^2\|\beta^*\|_1^2}{\rho^4}\left(\frac{1}{(n\wedge p)^{1-\frac{K}{2\zeta}}}\right)\log^5(np)$$

comes due to the pre-intervention error and the it follows immediately from Corollary 3.4. The second term,

$$\frac{C'''k^{5/2}}{\sqrt{n}}\|\beta^*\|_1$$

comes due to the generalization error of RSC/PCR for the post-intervention period. The proof of this bound on the generalization error of RSC/PCR follows in an identical fashion to Theorem 3.2 – the only change in the proof of Theorem 3.2 is that wherever an expectation over Ω was taken, we appropriately substitute it by taking an expectation over Θ, the latent distribution from which θ_i is sampled. □