Journal of Machine Learning Research 23 (2022) 1-60 Submitted 2/20; Revised 3/21; Published 2/22

Interpolating Predictors in High-Dimensional Factor Regression

Florentina Bunea FB238@CORNELL.EDU

Seth Strimas-Mackey* SCS324@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

Marten Wegkamp MHW73@CORNELL.EDU
Department of Mathematics and Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

Editor: Jon McAuliffe

Abstract

This work studies finite-sample properties of the risk of the minimum-norm interpolating predictor in high-dimensional regression models. If the effective rank of the covariance matrix Σ of the p regression features is much larger than the sample size n, we show that the min-norm interpolating predictor is not desirable, as its risk approaches the risk of trivially predicting the response by 0. However, our detailed finite-sample analysis reveals, surprisingly, that this behavior is not present when the regression response and the features are *jointly* low-dimensional, following a widely used factor regression model. Within this popular model class, and when the effective rank of Σ is smaller than n, while still allowing for $p \gg n$, both the bias and the variance terms of the excess risk can be controlled, and the risk of the minimum-norm interpolating predictor approaches optimal benchmarks. Moreover, through a detailed analysis of the bias term, we exhibit model classes under which our upper bound on the excess risk approaches zero, while the corresponding upper bound in the recent work Bartlett et al. (2020) diverges. Furthermore, we show that the minimum-norm interpolating predictor analyzed under the factor regression model, despite being model-agnostic and devoid of tuning parameters, can have similar risk to predictors based on principal components regression and ridge regression, and can improve over LASSO based predictors, in the high-dimensional regime.

Keywords: Interpolation, minimum-norm predictor, finite sample risk bounds, prediction, factor models, high-dimensional regression

1. Introduction

Motivated by the widely observed phenomenon that interpolating deep neural networks generalize well despite having zero training error, there has been a recent wave of literature showing that this is a general behaviour that can occur for a variety of models and prediction methods (Hastie et al., 2019; Feldman, 2019; Muthukumar et al., 2019; Mei and Montanari,

*. Corresponding author.

2019; Belkin et al., 2019a, 2018b,a, 2019b, 2018c; Jun et al., 2019; Mitra, 2019; Ma et al., 2017; Liang and Rakhlin, 2018; Xing et al., 2018; Bartlett et al., 2020).

One of the simplest settings is the prediction of a real-valued response $y \in \mathbb{R}$ from vector-valued features $X \in \mathbb{R}^p$ via generalized least squares (GLS). The GLS estimator $\widehat{\alpha} = \mathbf{X}^+\mathbf{y}$ is based on the Moore-Penrose pseudo-inverse of the $n \times p$ data matrix $\mathbf{X}$ and response vector $\mathbf{y} \in \mathbb{R}^n$, obtained from n i.i.d. copies (X_i, y_i), $i \in [n]$, of (X, y), with $p > n$. It coincides with the minimum-norm estimator, which in the case that $\mathbf{X}$ has full rank, interpolates the data. The interpolation property of $\widehat{\alpha}$ means that $\mathbf{X}\widehat{\alpha} = \mathbf{y}$. We refer to the corresponding predictor as the minimum-norm interpolating predictor.

This paper is devoted to the finite-sample statistical analysis of prediction via the generalized least squares estimator $\widehat{\alpha}$. We first note that ideally, the prediction risk $R(\widehat{\alpha}) := \mathbb{E}_{X,y}\left[(X^\top\widehat{\alpha} - y)^2\right]$ of $\widehat{\alpha}$ approaches the optimal risk $\inf_{\alpha\in\mathbb{R}^p} \mathbb{E}_{X,y}\left[(X^\top\alpha - y)^2\right]$. Unfortunately, this often turns out not to be the case. Theorem 1, stated in Section 2, proves that the ratio $R(\widehat{\alpha})/R(\mathbf{0})$ approaches 1 in the regime $\mathrm{r_e}(\Sigma_X) \gg n$. Clearly, this is undesirable as $R(\mathbf{0})$ is the non-optimal null risk of trivially predicting via the zero weight vector, ignoring the data. The *effective rank* $\mathrm{r_e}(\Sigma_X)$ of the $p \times p$ covariance matrix Σ_X of X is defined as the ratio between the trace of Σ_X and its operator norm, and is at most equal to its rank, $\mathrm{r_e}(\Sigma_X) \leq p$. In particular, if Σ_X is well-conditioned, with $\mathrm{r_e}(\Sigma_X) \asymp p$, then the prediction risk $R(\widehat{\alpha})$ of the minimum norm interpolator approaches the trivial risk $R(\mathbf{0})$, whenever $p \gg n$. This was previously observed, from a different perspective, in Hastie et al. (2019).

This opens the question as to whether, in the high-dimensional $p > n$ setting, there exist underlying distributions of the data that allow $R(\widehat{\alpha})$ to be close to an optimal risk benchmark. The recent work Bartlett et al. (2020) provides a positive answer to this question, primarily focusing on sufficient conditions on the spectrum of Σ_X that can lead to consistent prediction.

In this paper we show that the *joint* structure of (X, y), not just the marginal structure of X as considered in Bartlett et al. (2020), is important to understanding the conditions under which consistent prediction is possible with $\widehat{\alpha}$. In particular, we provide a detailed and novel finite-sample analysis of the prediction risk $R(\widehat{\alpha})$ when the pair (X, y) follows a linear factor regression model, $y = Z^\top\beta + \varepsilon$, $X = AZ + E$, in the regime

$$p \gg n \quad \text{but} \quad \mathrm{r_e}(\Sigma_X) < c \cdot n,$$

for an absolute constant $c > 0$. Here $(X, y) \in \mathbb{R}^p \times \mathbb{R}$ are observable random features and response, $Z \in \mathbb{R}^K$ is a vector of unobservable sub-Gaussian random latent factors with $K < p$, $A \in \mathbb{R}^{p\times K}$ is a loading matrix relating Z to X, and E and ε are mean-zero sub-Gaussian noise terms independent of Z and each other. Under this model, the observation made in inequality (7) of Section 3.1 below shows that $\mathrm{r_e}(\Sigma_X)$ is less than $c \cdot n$ as long as $K < c_1 \cdot n$ and the signal-to-noise ratio $\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| \gtrsim p/n \geq c_2 \cdot \mathrm{r_e}(\Sigma_E)/n$ for suitable absolute constants $c_1, c_2 > 0$. Here Σ_Z and Σ_E denote the covariance matrices of Z and E respectively, and ξ is the ratio between the Kth eigenvalue of $A\Sigma_Z A^\top$ and the operator norm of Σ_E. Section 3 is dedicated to deriving population-level properties of the factor regression model that are relevant to the performance of the GLS $\widehat{\alpha}$.

Our primary contribution is the study of $R(\widehat{\alpha})$ under the factor regression model, and in this regime. In Section 4 we present a detailed finite-sample study of the risk $R(\widehat{\alpha})$ of the model-agnostic interpolating predictor $\widehat{y}_x = X^\top\widehat{\alpha}$ in factor regression models with $p > n$ and

$K < n$, but with K allowed to grow with n. Our main result is Theorem 16 in Section 4.2. It provides a finite-sample bound on the *excess risk* $R(\widehat{\alpha}) - \sigma_\varepsilon^2$ of $\widehat{\alpha}$ in the high-dimensional setting $p > n$, relative to the natural risk benchmark $\mathbb{E}[\varepsilon^2] := \sigma_\varepsilon^2$ in the factor regression model; the excess risk relative to the benchmark $\inf_{\alpha\in\mathbb{R}^p} \mathbb{E}_{X,y}\left[(X^\top\alpha - y)^2\right]$ is also derived in this theorem. As a consequence, we obtain sufficient conditions under which the prediction risk $R(\widehat{\alpha})$ approaches the optimal risk, by adapting to the embedded dimension K. The excess risk not only decreases beyond the interpolation boundary to a non-zero value as observed in Hastie et al. (2019), but does indeed decrease to zero, as desired. We remark that at least for Gaussian (X, y), Bartlett et al. (2020) provides an alternative bound to Theorem 16. However, Theorem 16 provides an improved rate for typical factor regression models, and in particular provides examples when the upper bound on the excess risk in Bartlett et al. (2020) diverges, yet our results show that prediction is consistent; see Section 4.3 for a detailed comparison.

Table 1 below offers a snap-shot of our main results. The first row is a reminder that all results are established for $p > n$, while the second row separates the regimes of $r_e(\Sigma_X)$ larger or smaller than n. The third row specifies the assumptions on (X, y), namely sub-Gaussianity or, in addition, the factor regression model. The last row gives finite-sample bounds. The risk bounds in the bottom right panel are stated under the assumptions that the operator norms $\|\Sigma_Z\|$ and $\|\Sigma_E\|$ are constant and $\mathrm{r_e}(\Sigma_E) \asymp p$. These simplifying assumptions are made here for transparency of presentation and are not made in the body of the paper. The bottom right panel shows that the variance term V decreases if $p \gg n \log n$

$p > n$	
$\mathrm{r_e}(\Sigma_X) > C \cdot n$	$\mathrm{r_e}(\Sigma_X) < c \cdot n, \quad K < n$
(X, y) sub-Gaussian	(X, y) sub-Gaussian $y = \beta^\top Z + \varepsilon$ $X = AZ + E$
$\left\lvert \frac{R(\widehat{\alpha})}{R(\mathbf{0})} - 1 \right\rvert \lesssim \sqrt{n/\mathrm{r_e}(\Sigma_X)}$	$R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim B_Z + V$ $B_Z = \|\beta\|^2 \cdot p/(n \cdot \xi)$ $V = \{(n/p) + (K/n)\} \log n$

Table 1: Behavior of risk $R(\widehat{\alpha})$. Here $C > 1, c > 0$ are absolute constants with $C > c$. (i) $R(\widehat{\alpha})$ approaches null risk $R(\mathbf{0})$ for well-conditioned matrices Σ_X when $p \gg n$ (left panel); (ii) Variance term vanishes when $p \gg n \log n$ and $K \log n \ll n$; Bias term vanishes for $\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| \gg \|\beta\|^2 p/n$ (right panel).

and $K \log n \ll n$ and that the bias term B_Z decreases provided that the signal-to-noise ratio $\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\|$ is large enough. Specifically, we need that $\xi \gg \|\beta\|^2 p/n$, which for $\|\beta\|^2 \lesssim K$ amounts to $\xi \gg p \cdot K/n$. For instance, as explained in Section 3.1, a common, natural situation is $\xi \asymp p$ and the bias is small for $K \ll n$. In clustering problems where

the p coordinates of X can be clustered in K groups of approximately eqal size $m \approx p/K$ as discussed in Section 3.1, we find $\xi \asymp p/K$. In that case, B_Z vanishes if $n \gg K^2$.

We emphasize that a condition on the effective rank of Σ_X alone is not enough to guarantee that $R(\widehat{\alpha})$ is close to the optimal risk σ_ε^2. As argued in Section 3.4, if we assume the model $X = AZ + E$, but instead of assuming that y is also a function of Z, as in this work, we have a standard linear model $y = X^\top\theta + \eta$, with $\theta \in \mathbb{R}^p$, then the bias term *cannot* be ignored, unless $\|\theta\| \to 0$, which is typically not the case in high dimensions. In Section 3.3 we show that the best linear predictor $\alpha^* = \Sigma_X^+\Sigma_{Xy}$, that minimizes the risk $\mathbb{E}_{X,y}\left[(X^\top\alpha - y)^2\right]$, does in fact satisfy $\|\alpha^*\| \to 0$ under the factor regression model $y = Z^\top\beta + \varepsilon$ and thus that this is a natural setting for studying when the GLS generalizes well. From this perspective, this work illustrates the critical role played in the risk analysis by a modeling assumption in which (X, y) are jointly low-dimensional.

Finally, we remark that prediction under factor regression models has been well studied, starting with classical factor analysis that can be traced back to the 1940s (Joreskog, 1967, 1969, 1970, 1977; Lawley, 1940, 1941, 1943), including the pertinent work Anderson and Rubin (1956). A number of works ranging from purely Bayesian (Aguilar and West, 2000; Bhattacharya and Dunson, 2011; Hahn et al., 2013; Carvalho et al., 2008) to variational Bayes (Blei et al., 2017) to frequentist (Bing et al., 2019; Fan et al., 2013a, 2011, 2013b, 2017; Izenman, 2008; Jolliffe, 1982; Stock and Watson, 2002a,b, 2012) show that this class of models can be a useful framework for constructing and analyzing predictors of y from high-dimensional and correlated data. The literature on finite-sample prediction bounds under factor regression models is relatively limited, with instances provided by Bing et al. (2019); Fan et al. (2013a, 2011, 2013b, 2017), and most existing results established for K fixed. Relevant for the work presented here, the (non-Bayesian) prediction schemes that have been studied in generic factor regression models are often variations of principal component regression in $K < n$ fixed dimensions, and therefore typically do not interpolate the data. From this perspective, the results of this paper complement this existing literature, by studying the behavior of interpolating predictors in factor regression. Furthermore, in Section 4.4 we derive an upper bound on the excess risk of prediction based on principal components, under the factor regression model, and find that it is comparable to the excess risk bound of the interpolating predictor, in the regime $p \gg n$, provided that the covariance matrix Σ_E of the noise is well conditioned. This provides further motivation for the use of $\widehat{\alpha}$ in the setting discussed here.

The rest of the paper is organized as follows.

Section 2 derives sufficient conditions on Σ_X and $\sigma_y^2 := \mathbb{E}[y^2]$ under which $R(\widehat{\alpha})$ approaches the trivial risk $R(\mathbf{0})$. This section motivates the remainder of the paper, in which we study the risk behaviour when these conditions are violated.

Section 3 introduces the factor regression model (5) and derives population-level properties that are relevant to the performance of the GLS $\widehat{\alpha}$. Bounds on the effective rank and spectrum of Σ_X under (5) are given in Section 3.1, and reveal what key quantities to control in order to obtain non-trivial prediction risk bounds associated with the GLS estimate $\widehat{\alpha}$. Target risk benchmarks then are introduced in Section 3.2.

Section 3.3 investigates at the population level the properties of the best linear predictor $\alpha^* = \Sigma_X^+\Sigma_{Xy}$, under the factor regression model. We demonstrate the interesting

phenomenon that under model (5), $\|\alpha^*\| \to 0$ and yet $R(\alpha^*)/R(\mathbf{0}) \not\to 1$. We argue that this is in contrast to the behaviour of the best linear predictor θ in a standard linear regression model in which $\mathbb{E}[y|X] = X^\top\theta$ and typically $\|\theta\|$ is fixed or growing with p. We give a comparison between factor regression and standard linear regression in Section 3.4, commenting on assumptions on the operator norm of Σ_X, and on implications for prediction with the GLS.

The remainder of the paper, Section 4, contains our analysis of the GLS $\widehat{\alpha}$ and its prediction risk, under the factor regression model. Section 4.1 gives a preview of our main findings. In the noiseless case $\Sigma_E = 0$, we have that $\|\widehat{\alpha}\| \to 0$ (just like $\|\alpha^*\| \to 0$), but $R(\widehat{\alpha}) - R(\alpha^*)$ achieves the parametric rate K/n, up to a $\log(n)$ factor. In fact, we establish $X^\top\widehat{\alpha} = Z^\top\widehat{\beta}$ for the least squares estimate $\widehat{\beta}$ based on observed $(\mathbf{Z}, \mathbf{y})$.

Section 4.2 contains our main results in the more realistic setting $\Sigma_E \neq 0$. It establishes when $\widehat{\alpha}$ interpolates, and shows that typically $\|\widehat{\alpha}\| \to 0$, as in the noiseless case. Furthermore, in agreement with the findings in Section 4.1, $R(\widehat{\alpha})/R(\mathbf{0})$ does not approach 1. Instead, the finite-sample risk bound in Theorem 16 shows that under appropriate conditions on $\mathrm{r_e}(\Sigma_E)$ and the signal-to-noise ratio ξ, the excess risk $R(\widehat{\alpha}) - R(\alpha^*)$ converges to zero.

Section 4.3 presents a comparison with recent related work. In particular, we give a detailed comparison with Bartlett et al. (2020), which provides risk bounds for $\widehat{y}_x = X^\top\widehat{\alpha}$, for sub-Gaussian data (X, y), and offers sufficient conditions on Σ_X for optimal risk behavior, with emphasis on the optimality of the variance component of the risk. We present simplified versions of the generic bias and variance bounds obtained in Bartlett et al. (2020) under the factor regression model, which are derived in Appendix C.4. Table 2 of Section 4.3 summarizes our findings that the bound on the excess risk in Bartlett et al. (2020) is often larger in order of magnitude than the bound given in Theorem 16 of Section 4.2. In particular, we exhibit instances of the factor regression model class under which the excess risk upper bound in Bartlett et al. (2020) diverges, yet our upper bound approaches zero. We also compare our work to Mei and Montanari (2019), which gives an asymptotic analysis of the ridge regression estimator with arbitrarily small (but non-zero) regularization for a type of factor regression model.

Section 4.4 is devoted to a comparison with prediction via principal component regression and ℓ_1 and ℓ_2 penalized least squares, under the factor regression model.

All proofs and ancillary results are deferred to the Appendix. In particular, Theorem 30 in the Appendix complements Theorem 16 by showing the risk behavior of $\widehat{\alpha}$ for $n > c \cdot p$ for an absolute constant $c > 0$, and is included for completeness.

1.1 Notation

Throughout the paper, for a vector $v \in \mathbb{R}^d$, $\|v\|$ denotes the Euclidean norm of v.
For any matrix $A \in \mathbb{R}^{n\times m}$, $\|A\|$ denotes the operator norm and A^+ the Moore-Penrose pseudo-inverse. See Appendix E for a definition of the pseudo-inverse and a summary of its properties used in this paper.
For a positive semi-definite matrix $Q \in \mathbb{R}^{p\times p}$, and vector $v \in \mathbb{R}^p$, we define $\|v\|_Q^2 := v^\top Q v$, let $\lambda_1(Q) \geq \lambda_2(Q) \geq \cdots \geq \lambda_p(Q)$ be its ordered eigenvalues, $\kappa(Q) := \lambda_1(Q)/\lambda_p(Q)$ its condition number, and $\mathrm{r_e}(Q) := \mathrm{tr}(Q)/\|Q\|$ its effective rank.

The identity matrix in dimension m is denoted I_m.
The set $\{1, 2, \ldots, m\}$ is denoted $[m]$.
Letters c, c', c_1, C, etc., are used to denote absolute constants, and may change from line to line.

2. Interpolation and the Null Risk

Given i.i.d. observations $(X_1, y_1), \ldots, (X_n, y_n)$, distributed as $(X, y) \in \mathbb{R}^p \times \mathbb{R}$, let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the corresponding data matrix with rows $X_1, \ldots X_n$, and let $\mathbf{y} \coloneqq (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. For the rest of the paper, unless specified otherwise, we make the blanket assumption that $p > n$.

We are interested in studying the prediction risk associated with the minimum ℓ_2-norm estimator $\widehat{\alpha}$ defined as

$$\widehat{\alpha} \coloneqq \arg\min \left\{ \|\alpha\| : \ \|\mathbf{X}\alpha - \mathbf{y}\| = \min_u \|\mathbf{X}u - \mathbf{y}\| \right\}. \tag{1}$$

We define the prediction risk for any $\alpha \in \mathbb{R}^p$ as

$$R(\alpha) \coloneqq \mathbb{E}_{X,y}[(X^\top \alpha - y)^2]. \tag{2}$$

The expectation is over the new data point (X, y), independent of the observed data $(\mathbf{X}, \mathbf{y})$. In particular, since $\widehat{\alpha}$ is independent of (X, y), we have $R(\widehat{\alpha}) = \mathbb{E}_{X,y}\left[(X^\top \widehat{\alpha} - y)^2 \,|\, \mathbf{X}, \mathbf{y}\right] = \mathbb{E}_{X,y}\left[(X^\top \widehat{\alpha} - y)^2\right]$. If the data matrix $\mathbf{X}$ has full rank of $n < p$, then $\min_{u \in \mathbb{R}^p} \|\mathbf{X}u - \mathbf{y}\| = 0$ and

$$\widehat{\alpha} \coloneqq \arg \min_{\alpha:\ \mathbf{X}\alpha = \mathbf{y}} \|\alpha\|. \tag{3}$$

Regardless of the rank of $\mathbf{X}$, Equation (1) always has the closed form solution $\widehat{\alpha} = \mathbf{X}^+ \mathbf{y}$, where $\mathbf{X}^+$ is the Moore-Penrose pseudo-inverse of $\mathbf{X}$; we prove this fact in section D.1 for completeness. We begin our consideration of the minimum-norm estimator $\widehat{\alpha} = \mathbf{X}^+ \mathbf{y}$ by showing that its risk $R(\widehat{\alpha})$ approaches the null risk $R(\mathbf{0})$ whenever the effective rank $\mathrm{r_e}(\Sigma_X)$ grows at a rate faster than n. Proofs for this section are contained in Appendix A. We make the following distributional assumption.

Assumption 1. *$X = \Sigma_X^{1/2} \tilde{X}$ and $y = \sigma_y \tilde{y}$, where $\tilde{X} \in \mathbb{R}^p$ has independent entries, and both $\tilde{X}$ and $\tilde{y}$ have zero mean, unit variance, and sub-Gaussian constants bounded by an absolute constant.*

Theorem 1. *Suppose Assumption 1 holds and $\mathrm{r_e}(\Sigma_X) > C \cdot n$ for some absolute constant $C > 1$ large enough. Then, with probability at least $1 - c\mathrm{e}^{-c'n}$ for absolute constants $c, c' > 0$,*

$$\left| \frac{R(\widehat{\alpha})}{R(\mathbf{0})} - 1 \right| \lesssim \sqrt{\frac{n}{\mathrm{r_e}(\Sigma_X)}}. \tag{4}$$

As a consequence, $\widehat{\alpha}$ is not a useful estimator in the regime $r_e(\Sigma_X) \gg n$, as trivially predicting with the null vector $\mathbf{0} \in \mathbb{R}^p$ will give asymptotically equivalent results. This occurs, for instance, when Σ_X is well conditioned and $p/n \to \infty$. Figure 2 in Hastie et al. (2019) depicts an example of this behavior: it plots $\mathbb{E}[\|\widehat{\alpha} - \alpha\|^2 | \mathbf{X}]$ as a function of the ratio $\gamma = p/n$, where (X, y) follows the linear model $y = \alpha^\top X + \varepsilon$ with $\Sigma_X = I_p$.

This motivates the study of $R(\widehat{\alpha})$ when the condition $\mathrm{r_e}(\Sigma_X) > C \cdot n$ of Theorem 1 fails. The recent work Bartlett et al. (2020) developed bounds for the excess risk $R(\widehat{\alpha}) - \inf_{\alpha \in \mathbb{R}^p} R(\alpha)$ under the linearity assumption $\mathbb{E}[y|X] = X^\top \theta$ (for some $\theta \in \mathbb{R}^p$), and used this to show that the excess risk goes to zero for a certain class of *benign* covariance matrices that in particular satisfy $\mathrm{r_e}(\Sigma_X)/n \to 0$ and $\|\Sigma_X\| = 1$.

In this work we are interested in obtaining risk bounds for $R(\widehat{\alpha})$ under a different model, the factor regression model (5) given below. In this model, while $\mathrm{r_e}(\Sigma_X)/n$ remains bounded, $\|\Sigma_X\|$ typically grows with p (see Lemma 4 below), in contrast to the assumption $\|\Sigma_X\| = 1$ of the definition of benign matrices in Bartlett et al. (2020). Furthermore, the results in Bartlett et al. (2020) only apply to model (5) when (X, y) are assumed to be jointly Gaussian. In this case, their bound offers an alternative result, which we compare to our main result in Section 4.3 below. We find that in this common regime, we obtain a tighter bound.

3. Factor Regression Models

In this paper, we consider the factor regression model (FRM). This is a latent factor model in which we single out one variable, $y \in \mathbb{R}$, to emphasize its role as the response relative to input covariates $X \in \mathbb{R}^p$, while both X and y are directly connected to a lower dimensional, unobserved, random vector $Z \in \mathbb{R}^K$, with mean zero and $K < n$. Specifically, the factor regression model postulates that

$$X = AZ + E, \qquad y = Z^\top \beta + \varepsilon, \tag{5}$$

where $\beta \in \mathbb{R}^K$ is the latent variable regression vector, $A \in \mathbb{R}^{p \times K}$ is a unknown loading matrix, and $\varepsilon \in \mathbb{R}$ and $E \in \mathbb{R}^p$ are mean zero additive noise terms independent of one another and of Z. We let $\Sigma_E \coloneqq \mathrm{Cov}(E)$, $\Sigma_Z \coloneqq \mathrm{Cov}(Z)$ and $\sigma_\varepsilon^2 \coloneqq \mathrm{Var}(\varepsilon)$. For the remainder of the paper we will assume that the data consist of n i.i.d. pairs (X_i, y_i) satisfying (5), in that

$$X_i = AZ_i + E_i, \qquad y_i = Z_i^\top \beta + \varepsilon_i \qquad \forall i \in [n], \tag{6}$$

where the latent factors $Z_1, \ldots, Z_n \in \mathbb{R}^K$ are i.i.d. copies of Z, and the error terms $E_i \in \mathbb{R}^p$ and $\varepsilon_i \in \mathbb{R}$ for $i = 1, \ldots, n$ are i.i.d. copies of E and ε, respectively. We recall that $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the matrix with rows $X_1, \ldots, X_n$ and $\mathbf{y} \in \mathbb{R}^n$ is the vector with entries $y_1, \ldots, y_n$. We similarly let $\mathbf{Z} \in \mathbb{R}^{n \times K}$ be the matrix with rows $Z_1, \ldots, Z_n$.

The remainder of this section is dedicated to deriving population-level properties of the factor regression model that are relevant to the performance of the GLS $\widehat{\alpha}$. In particular, we will (1) bound the effective rank of Σ_X, (2) bound the eigenvalues of Σ_X, (3) define two natural risk benchmarks and show when they are asymptotically equivalent, (4) show that the weight vector of the best linear predictor has vanishing norm, and (5) prove that, nonetheless, the null risk $R(\mathbf{0})$ is clearly sub-optimal. The first two properties reflect the low-rank structure of the covariance matrix Σ_X and are presented in Section 3.1. The risk benchmarks are introduced and analyzed in Section 3.2. Section 3.3 investigates the properties of the best linear predictor $\alpha^* = \Sigma_X^+ \Sigma_{Xy}$ at the population level, showing properties (4) and (5). The fourth property in particular is a consequence of the joint low-dimensional structure of (X, y) via the vector of covariances Σ_{Xy}. It is a distinct property of the factor

regression model that sets it apart from the classical regression model where the response y is linearly related to X via $\mathbb{E}[y|X] = \theta^\top X$. We present a comparison between factor regression and classical linear regression in Section 3.4.

3.1 Effective Rank and Spectrum of Σ_X in the FRM

Theorem 1 and its discussion above imply that in order for the generalized least squares estimator $\widehat{\alpha}$ to have asymptotically better prediction performance than the trivial estimator $\mathbf{0} \in \mathbb{R}^p$, the ratio $\mathrm{r_e}(\Sigma_X)/n$ must remain bounded as n and p grow, as a first requirement.

Using that $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$ under (5), we find

$$\begin{aligned}
\mathrm{r_e}(\Sigma_X) &= \frac{\mathrm{tr}(\Sigma_X)}{\|\Sigma_X\|} \\
&\leq \frac{\mathrm{tr}(A\Sigma_Z A^\top) + \mathrm{tr}(\Sigma_E)}{\|A\Sigma_Z A^\top\|} && (\text{since } \|\Sigma_X\| \geq \|A\Sigma_Z A^\top\|) \\
&\leq K + \frac{\mathrm{tr}(\Sigma_E)}{\|A\Sigma_Z A^\top\|} && (\text{since } \mathrm{tr}(A\Sigma_Z A^\top) \leq K\|A\Sigma_Z A^\top\|) \\
&\leq K + \frac{\|\Sigma_E\|}{\lambda_K(A\Sigma_Z A^\top)} \cdot \frac{\mathrm{tr}(\Sigma_E)}{\|\Sigma_E\|}, && (\text{since } \|A\Sigma_Z A^\top\| \geq \lambda_K(A\Sigma_Z A^\top))
\end{aligned}$$

where we use the convention that $\mathrm{tr}(\Sigma_E)/\|\Sigma_E\| = \mathrm{r_e}(\Sigma_E) = 1$ if $\Sigma_E = 0$. We thus have

$$\frac{\mathrm{r_e}(\Sigma_X)}{n} \leq \frac{K}{n} + \frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}, \tag{7}$$

where

$$\xi \coloneqq \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\|, \tag{8}$$

can be viewed as a signal-to-noise ratio since $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$, and we use the convention that $\xi = \infty$ and $\mathrm{r_e}(\Sigma_E)/\xi = 0$ when $\Sigma_E = 0$. In standard factor regression models (Anderson and Rubin, 1956), $\Sigma_E = I_p$, in which case $\mathrm{r_e}(\Sigma_E) = p$, but in our analysis we allow for a general Σ_E, with possibly smaller $\mathrm{r_e}(\Sigma_E)$. The following simple result follows directly from (7).

Lemma 2. *Under model (5), we have* $\mathrm{r_e}(\Sigma_X)/n \leq c_3$ *whenever*

$$\frac{K}{n} \leq c_1 \quad \text{and} \quad \xi \geq c_2\frac{\mathrm{r_e}(\Sigma_E)}{n}, \tag{9}$$

for positive absolute constants c_1, c_2, c_3.

Remark 3. *We remark on conditions under which (9) holds. Suppose that the eigenvalues of* Σ_Z *and* Σ_E *are constant, that is,* $c_1 \leq \lambda_K(\Sigma_Z) \leq \|\Sigma_Z\| \leq C_1$ *and* $c_2 < \lambda_p(\Sigma_E) \leq \|\Sigma_E\| < C_2$, *for some* $c_1, c_2, C_1, C_2 \in (0, \infty)$, *both standard assumptions in factor models. Then,*

$$\mathrm{r_e}(\Sigma_E) \asymp p, \quad \text{and} \quad \xi = \frac{\lambda_K(A\Sigma_Z A^\top)}{\|\Sigma_E\|} \asymp \lambda_K(A^\top A), \tag{10}$$

so the condition (9) reduces to $K/n \leq c_1$ *and*

$$\lambda_K(A^\top A) \gtrsim \frac{p}{n}. \tag{11}$$

We give a few examples of A *that imply (11):*

1. *For a well-conditioned matrix $A \in \mathbb{R}^{p\times K}$ with entries taking values in a bounded interval, $\lambda_K(A^\top A) \asymp p$, and (11) holds.*

2. *Treating A as a realization of a random matrix with i.i.d. entries and $p \gg K$, then by standard concentration arguments (see Vershynin (2019), for example) we once again have $\lambda_K(A^\top A) \gtrsim p$, with high probability, and (11) holds.*

3. *In other situations, (11) is an assumption. It is a very natural, and mild, requirement in factor regression models, and if A is structured and sparse, (11) can be given further interpretation. For instance, the model $X = AZ + E$ has been used and analyzed in Bunea et al. (2019) for clustering the p components of X around the latent Z-coordinates, via an assignment matrix $A \in \{0,1\}^{p\times K}$, and when Σ_E is an approximately diagonal matrix. Denoting the size of the smallest of the K non-overlapping clusters by m, for some integer $2 \le m \le p$, it is immediate to see (Lemma 31 in Appendix D.4) that $\lambda_K(A^\top A) \ge m$. Furthermore, when these K clusters are approximately balanced, then $m \approx p/K$ and (11) holds, provided $K \lesssim n$.*

The positive repercussion of Lemma 2 is that under condition (9) and for small enough constant c_3, Theorem 1 no longer applies. This in turn opens up the possibility of showing that, under the data generating model (5) with restrictions (9), the risk $R(\widehat{\alpha})$ will approach optimal risk benchmarks. We define the benchmark risks in terms of the best linear predictors of y from X and Z, respectively, in Section 3.2, and show that $R(\widehat{\alpha})$ can indeed approach these benchmarks in Sections 4.1 and 4.2.

For completeness, we offer the following result characterizing the spectrum of Σ_X under the factor regression model. In particular, as announced in Section 2, we find that the operator norm $\|\Sigma_X\|$ diverges with p under mild conditions. The proof can be found in Appendix B.1.

Lemma 4. *Suppose that for some $c_1, c_2, C_1, C_2 \in (0,\infty)$,*

$$c_1 \le \lambda_K(\Sigma_Z) \le \|\Sigma_Z\| \le C_1 \quad \text{and} \quad c_2 < \lambda_p(\Sigma_E) \le \|\Sigma_E\| < C_2. \tag{12}$$

The spectrum of Σ_X can then be characterized as follows:

1. *$\lambda_i(\Sigma_X) \ge c_2 > 0$ for all $i \in [p]$, i.e., the entire spectrum of Σ_X is bounded below;*

2. *$\lambda_K(\Sigma_X) \ge c_1\lambda_K(A^\top A)$, so the first K eigenvalues of Σ_X diverge if $\lambda_K(A^\top A) \to \infty$ as $p \to \infty$;*

3. *$c_2 \le \lambda_i(\Sigma_X) \le C_2$ for $i > K$, i.e., the last $p-K$ eigenvalues of Σ_X are bounded above and below.*

After introducing the risk benchmarks below, we investigate the behaviour of the best linear prediction vector $\alpha^* = \Sigma_X^+\Sigma_{Xy}$ of y from X under the factor regression model in Section 3.3, and use this in Section 3.4 to clarify the importance of the factor regression model, in which (X, y) *jointly* have a low-dimensional structure, in contrast to the classical linear model $y = X^\top\theta + \eta$ with low-dimensional structure on X alone.

3.2 Risk Benchmarks

We introduce here two natural benchmarks for $R(\widehat{\alpha})$ under the factor regression model, and characterize their relationship. Under model (5), if $Z \in \mathbb{R}^K$ were observed, the optimal risk of a linear oracle with access to Z is

$$\min_{v \in \mathbb{R}^K} \mathbb{E}\left[(Z^\top v - y)^2\right] = \mathbb{E}[\varepsilon^2] = \sigma_\varepsilon^2, \tag{13}$$

which we henceforth refer to as the oracle risk. Another natural benchmark to compare the risk $R(\widehat{\alpha})$ to is the minimum risk possible for any linear predictor $\alpha^\top X$, namely $R(\alpha^*)$, where

$$\alpha^* \in \arg\min_{\alpha \in \mathbb{R}^p} R(\alpha). \tag{14}$$

Lemma 27 in Appendix D shows that for arbitrary zero-mean (X, y) with finite second moments, $\alpha^* = \Sigma_X^+ \Sigma_{Xy}$ is a minimizer of $R(\alpha)$, where $\Sigma_{Xy} \coloneqq \mathbb{E}[Xy] \in \mathbb{R}^p$ is the vector of component-wise covariances.

We can characterize the difference between these two benchmarks, σ_ε^2 and $R(\alpha^*)$, as follows. See Appendix B.2 for the proof of this result.

Lemma 5 (Comparison of risk benchmarks)**.** *Suppose model (5) holds and let ξ be the signal-to-noise ratio defined in (8). We have*

1. *$R(\alpha^*) - \sigma_\varepsilon^2 \geq 0$ with equality if $\Sigma_E = 0$.*
2. *Provided the matrices Σ_Z, Σ_E, and A are full rank,*

$$\frac{\xi}{1+\xi} \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \leq R(\alpha^*) - \sigma_\varepsilon^2 \leq \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta,$$

where

$$\beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \leq \frac{1}{\xi} \|\beta\|_{\Sigma_Z}^2.$$

In particular, $\|\beta\|_{\Sigma_Z}^2 / \xi \to 0$ implies $R(\alpha^*) - \sigma_\varepsilon^2 \to 0$, as $p \to \infty$.

Although the optimal risk $R(\alpha^*)$ is always greater than the oracle risk σ_ε^2 (part 1 of Lemma 5), the bound $\|\beta\|_{\Sigma_Z}^2 / \xi$ on the difference $R(\alpha^*) - \sigma_\varepsilon^2$ in part 2 of Lemma 5 is not a leading term in the excess risk bound given in Theorem 16. From this perspective, we can view these benchmarks as asymptotically equivalent, but with different interpretations. Interestingly, the condition $\lim_{p\to\infty} \|\beta\|_{\Sigma_Z}^2 / \xi = 0$ forces $\|\alpha^*\| \to 0$, see Corollary 9 in the next section. This is an important feature of the FRM, and its repercussions are discussed in Section 3.4.

3.3 Best Linear Prediction in Factor Regression Models (Population Level)

In this section we investigate the properties of the population-level predictor α^*, defined in (14), under the factor regression model (5). In particular, we prove that $\|\alpha^*\| \to 0$ and yet $R(\mathbf{0}) - R(\alpha^*) > 0$ under the conditions

$$\lim_{p\to\infty} \|\beta\|_{\Sigma_Z}^2 / \lambda_K(A \Sigma_Z A^\top) = 0 \quad \text{and} \quad \liminf_{p\to\infty} \|\beta\|_{\Sigma_Z} > 0. \tag{15}$$

The property $\|\alpha^*\| \to 0$ in particular is a consequence of the joint low-dimensional structure of (X, y) via the covariance $\Sigma_{Xy} = A\Sigma_Z\beta$, which the vector $\alpha^* = \Sigma_X^+\Sigma_{Xy}$ depends on. Proofs for this section can be found in Appendix B.3. We first characterize the norms $\|\alpha^*\|$ and $\|\alpha^*\|_{\Sigma_X}$; the latter norm is of interest via the identity

$$R(\mathbf{0}) - R(\alpha^*) = \|\alpha^*\|_{\Sigma_X}^2. \tag{16}$$

It is instructive to first consider the simple case of noiseless features, $X = AZ$, with $E = 0$. In this case, the best linear predictor of y from X is ${\alpha^*}^\top X = (A^\top\alpha^*)^\top Z$. The following lemma states that $\alpha^* = A^{+\top}\beta$, which by the identity $A^\top A^{+\top} = I_K$ when A is full rank gives

$${\alpha^*}^\top X = (A^\top A^{+\top}\beta)^\top Z = \beta^\top Z, \tag{17}$$

showing that the best linear predictor from X reduces to the best linear predictor from Z. The lemma then uses this to derive explicit expressions for the norms of α^*.

Lemma 6. *Suppose model (5) holds, that $\Sigma_E = 0$, and that Σ_Z and A are full rank. Then, $\alpha^* = A^{+\top}\beta$, and*

$$\|\alpha^*\|_{\Sigma_X}^2 = \|\beta\|_{\Sigma_Z}^2 \quad \text{and} \quad \|\alpha^*\|^2 = \beta^\top(A^\top A)^{-1}\beta.$$

We next find that in the more realistic case, when $\Sigma_E \neq 0$, even though identity (17) no longer holds, we can recover the same identities for $\|\alpha^*\|_{\Sigma_X}$ and $\|\alpha^*\|$, up to constants, when the noise matrix Σ_E is well-conditioned.

Lemma 7. *Suppose model (5) holds and that A, Σ_Z, Σ_E are all full rank. Then, when $\xi = \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| > c > 1$ and $\kappa(\Sigma_E) < C < \infty$,*

$$\|\alpha^*\|_{\Sigma_X}^2 \asymp \|\beta\|_{\Sigma_Z}^2 \quad \text{and} \quad \|\alpha^*\|^2 \asymp \beta^\top(A^\top A)^{-1}\beta.$$

Remark 8. *We illustrate our findings in Lemmas 6 and 7 with the following example (that we will use in our simulations in Section 4.4), where $\Sigma_Z = \sigma_Z^2 I_K$, $\Sigma_E = \sigma_E^2 I_p$, and $A^\top A = a^2 I_K$. It can be verified that in this case,*

$$\alpha^* = \frac{\sigma_Z^2}{\sigma_E^2 + a^2\sigma_Z^2} A\beta \tag{18}$$

$$\|\alpha^*\|^2 = \frac{a^2\sigma_Z^2}{(\sigma_E^2 + a^2\sigma_Z^2)^2}\|\beta\|_{\Sigma_Z}^2 \tag{19}$$

$$\|\alpha^*\|_{\Sigma_X}^2 = \frac{a^2\sigma_Z^2}{\sigma_E^2 + a^2\sigma_Z^2}\|\beta\|_{\Sigma_Z}^2. \tag{20}$$

Since $\lambda_K(A\Sigma_Z A^\top) = a^2\sigma_Z^2$ and $\xi = a^2\sigma_Z^2/\sigma_E^2$, it confirms that $\|\beta\|_{\Sigma_Z}^2/\lambda_K(A\Sigma_Z A^\top) \to 0$ forces $\|\alpha^*\| \to 0$, while at the same time $\|\alpha^*\|_{\Sigma_X}^2 \asymp \|\beta\|_{\Sigma_Z}^2$ when ξ is bounded below (in fact, $\|\alpha^*\|_{\Sigma_X}^2/\|\beta\|_{\Sigma_Z}^2 \to 1$ when $\xi \to \infty$ in this example).

We note that while $\|\alpha^*\| \to 0$, there is no reason to assume α^* to be sparse. In this example, we can see from the explicit formula (18) that $\alpha_i^* = 0 \iff A_{i\bullet}^\top\beta = 0$, whence row-sparsity of the matrix A induces sparsity of the vector α^*. For a more general A, this isn't the case and α^* isn't necessarily sparse or even approximately sparse. This observation is corroborated in our simulations in Section 4.4.

Identity (16), Lemma 6 and Lemma 7 imply the following conclusion.

Corollary 9. *Suppose model (5) holds with A, Σ_Z, Σ_E all full rank, let $\xi = \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| > c > 1$, and suppose $\kappa(\Sigma_E) < C < \infty$. Alternatively, suppose that under model (5), $\Sigma_E = 0$ and A, Σ_Z are full rank. Then, in either case, condition (15) implies*

$$\lim_{p\to\infty} \|\alpha^*\| = 0, \quad \textit{while} \quad \liminf_{p\to\infty} \{R(\mathbf{0}) - R(\alpha^*)\} \gtrsim \liminf_{p\to\infty} \|\beta\|^2_{\Sigma_Z} > 0.$$

This result shows that while the norm of α^* converges to zero in the factor regression model, its risk is separated from the risk of the null predictor $\mathbf{0}$ by a constant times $\|\beta\|^2_{\Sigma_Z}$. In fact, as β is an arbitrary vector in $\mathbb{R}^K$, the gap $R(\mathbf{0}) - R(\alpha^*)$ will typically grow as K increases.

The behaviour $\|\alpha^*\| \to 0$ is a feature of the factor regression model that arises from the joint low-dimensional structure of the model, as encoded in the covariance Σ_{Xy}. This is in stark contrast to the behaviour of the best linear prediction vector θ in a linear model $y = X^\top\theta + \eta$, as we do not expect $\|\theta\|$ to vanish as p grows. We discuss the important roles played by these quantities in the risk bound analysis in the next section.

3.4 Prediction Under Linear Regression with Conditions on the Design Versus Prediction Under Latent Factor Regression

The model (5) can be said to have *joint* low-dimensional structure, in that both the features X and response y are (noisy) functions of the low-dimensional latent vector Z. We would like to argue that this structure plays an important role in the behaviour of the GLS $\widehat{\alpha}$, which we will study in the next section. In particular, to understand the implications of this joint-low dimensional structure, we could compare model (5) to a model in which X continues to follow a factor model, but y is connected to X via a linear model:

$$X = AZ + E, \qquad y = X^\top\theta + \eta, \tag{21}$$

where $\theta \in \mathbb{R}^p$ is a generic p-dimensional regression vector, and η is zero-mean noise independent of X. Model (21) captures the setting in which there is low-dimensional structure in the features alone.

When $(X, y) \in \mathbb{R}^p \times \mathbb{R}$ are jointly Gaussian, Lemma 29 in Appendix D.2 shows the simple fact that if the factor regression model (5) holds, then (21) holds, with regression coefficients $\theta = \alpha^*$ and error $\eta := y - X^\top\alpha^*$, independent of X. Here α^* is the best linear predictor *under the factor regression model (5)*, which we studied the properties of in Section 3.3 above.

We can thus compare model (5) and (21) directly in the Gaussian case. We stress that we do not assume Gaussianity elsewhere in our paper, but use it here to facilitate this comparison.

In Section 3.3 we found that $\|\alpha^*\| \to 0$, provided (15) holds. Thus, when the factor regression model (5) is viewed as a particular case of (21), we have $\|\alpha^*\| = \|\theta\| \to 0$. This behavior is in sharp contrast with the typical behavior of a generic linear model $y = X^\top\theta + \eta$ as in (21), in which $\|\theta\|$ is usually fixed or growing with p. We argue that this difference has important implications for the performance of the GLS predictor $\widehat{\alpha}$.

One way this can be seen is by considering the bound from the recent work Bartlett et al. (2020) on the excess risk $R(\widehat{\alpha}) - R(\theta)$, proved under model $E(y|X) = X^T\theta$ for sub-Gaussian (X, y). In particular, the bound of Bartlett et al. (2020) contains a bias term given by

$$\|\theta\|^2 \|\Sigma_X\| \max\left\{\sqrt{\frac{\mathrm{r_e}(\Sigma_X)}{n}}, \frac{\mathrm{r_e}(\Sigma_X)}{n}\right\}. \tag{22}$$

We examine this bound assuming further that model (21) holds. Since

$$\|\Sigma_X\| \max\left\{\sqrt{\frac{\mathrm{r_e}(\Sigma_X)}{n}}, \frac{\mathrm{r_e}(\Sigma_X)}{n}\right\} = \max\left\{\sqrt{\frac{\|\Sigma_X\|\mathrm{tr}(\Sigma_X)}{n}}, \frac{\mathrm{tr}(\Sigma_X)}{n}\right\} \geq \frac{\mathrm{tr}(\Sigma_X)}{n} \tag{23}$$

and

$$\frac{\mathrm{tr}(\Sigma_X)}{n} = \frac{\mathrm{tr}(\Sigma_E)}{n} + \frac{\mathrm{tr}(A\Sigma_Z A^\top)}{n} \to \infty$$

under model (21) with mild assumptions on either Σ_E (e.g., $\Sigma_E \asymp I_p$) or A (see Remark 3), the bias term (22) will only converge to zero if $\|\theta\| \to 0$.

As noted above, $\|\theta\| \to 0$ is rather unnatural in a generic model (21). However, we also noted that when (X, y) are Gaussian and the factor regression model (5) holds, then (21) holds with $\|\theta\| = \|\alpha^*\| \to 0$, which means that the bias term (22) can converge to zero when the data is generated by model (5). We take this as indication that the bias in prediction with $\widehat{\alpha}$ can be significantly lower in the factor regression model (5) compared to a generic model (21) as a result of the joint low-dimensional structure of model (5).

We note that this discussion is only based on an upper bound (22) on the bias term of the prediction risk. It nevertheless motivates a full investigation of an alternative upper bound to (22), directly derived under model (5). This is the subject of Section 4 below, with our main result presented in Theorem 16.

Remark 10. *The authors of Bartlett et al. (2020) take a different route, complementary to ours, in their analysis of the bound (22). Although they derived it with no assumptions on $\|\Sigma_X\|$, the desired convergence to zero is established under the assumption that Σ_X belongs to what is called in Bartlett et al. (2020) a class of benign covariance matrices, that in particular satisfy $\|\Sigma_X\| = 1$.*

This assumption allows the authors to avoid making the unpleasant assumption that a generic θ would have ℓ_2-norm converging to zero with p. To see why, note that when $\|\Sigma_X\|$ is bounded, working in the regime $\mathrm{r_e}(\Sigma_X)/n \to 0$ immediately implies

$$\|\Sigma_X\| \max\left\{\sqrt{\frac{\mathrm{r_e}(\Sigma_X)}{n}}, \frac{\mathrm{r_e}(\Sigma_X)}{n}\right\} \to 0,$$

which in turn means that under the assumption $\|\Sigma_X\| = 1$, their bias term (22) can converge to zero even when $\|\theta\| \not\to 0$, for a generic θ.

However, as we have shown in Lemma 4 above, this class does not cover covariance matrices Σ_X associated with a random vector that obeys a factor model $X = AZ + E$, as $\|\Sigma_X\| \to \infty$ with p in this case. Since in factor regression we argued that $\|\theta\| = \|\alpha^*\| \to 0$, one can still expect that (22) will vanish, in the regime $\mathrm{r_e}(\Sigma_X)/n \to 0$, even though $\|\Sigma_X\| \to \infty$. The results of Section 4 can thus be viewed as complementary to those in Bartlett et al. (2020).

4. Minimum ℓ_2-norm Prediction in Factor Regression

In this section we analyze the GLS $\widehat{\alpha}$, and present our main contribution, namely, novel finite-sample bounds on the prediction risk $R(\widehat{\alpha})$ relative to the benchmarks laid out in Section 3.2.

4.1 Exact Adaptation in Factor Regression Models with Noiseless Features

We begin our analysis by considering an extreme case of model (5), in which $E = 0$ almost surely, and thus Σ_X is degenerate, with $\mathrm{r_e}(\Sigma_X) \leq \mathrm{rank}(\Sigma_X) = K$.

Proofs for this section are contained in Appendix C.1. We make the following assumptions.

Assumption 2. *The $p \times K$ matrix A and $K \times K$ matrix Σ_Z both have full rank equal to K.*

Assumption 3. *$E = \Sigma_E^{1/2}\tilde{E}$, where $\tilde{E} \in \mathbb{R}^p$ has independent entries with zero mean, unit variance, and sub-Gaussian constants bounded by an absolute constant.*

Furthermore, $Z = \Sigma_Z^{1/2}\tilde{Z}$ and $\varepsilon = \sigma_\varepsilon\tilde{\varepsilon}$, where $\tilde{Z} \in \mathbb{R}^K$ and $\tilde{\varepsilon} \in \mathbb{R}$ have zero mean and sub-Gaussian constants bounded by an absolute constant.

We first analyze the norm of $\widehat{\alpha}$. In Lemma 6 above, we showed that $\|\alpha^*\|^2 = \beta^\top(A^\top A)^{-1}\beta$ when $\Sigma_E = 0$, and as a result, Corollary 9 states that $\|\alpha^*\| \to 0$, provided $\|\beta\|^2_{\Sigma_Z}/\lambda_K(A\Sigma_Z A^\top) \to 0$ as $p \to \infty$. We now show that $\widehat{\alpha}$ mimics this behavior under the additional condition that $(\sigma_\varepsilon^2 \log n)/\lambda_K(A\Sigma_Z A^\top) \to 0$ as $n \to \infty$.

Lemma 11. *Under model (5) with $\Sigma_E = 0$, suppose that Assumptions 2 and 3 hold, and that $n > C \cdot K$ for some large enough absolute constant $C > 0$. Then, with probability at least $1 - c/n$ for some absolute constant $c > 0$,*

$$\|\widehat{\alpha}\|^2 \lesssim \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\left(\|\beta\|^2_{\Sigma_Z} + \sigma_\varepsilon^2\frac{K\log n}{n}\right). \tag{24}$$

The fact that $\widehat{\alpha}$ vanishes does *not* imply that $R(\widehat{\alpha})/R(\mathbf{0}) \to 1$, just like $R(\alpha^*)/R(\mathbf{0}) \not\to 1$ in Corollary 9. We will now show that in fact the risk $R(\widehat{\alpha})$ approaches the optimal risk $R(\alpha^*)$ by adapting to the low-dimensional structure of the factor regression model. Let $\widehat{y}_z \coloneqq Z^\top\widehat{\beta}$ be the predictor based on the least-squares regression coefficients $\widehat{\beta} \coloneqq \mathbf{Z}^+\mathbf{y}$ of $\mathbf{y}$ onto $\mathbf{Z}$; this is the classical least-squares prediction of y under model (5) that an oracle would use if it had access to the unobserved data matrix $\mathbf{Z}$, and the new, but unobservable, data point Z. In contrast, let $\widehat{y}_x = X^\top\widehat{\alpha}$ be the least-squares predictor of y from X based on $(\mathbf{X}, \mathbf{y})$ only. Theorem 12.1 below shows that the realizable prediction $\widehat{y}_x$ equals the oracle prediction $\widehat{y}_z$. The second part of the theorem gives lower and upper bounds on the risk that hold with high probability over the training data.

Theorem 12 (Factor regression with noiseless features)**.** *Under model (5) with $\Sigma_E = 0$, suppose that Assumption 2 holds.*

1. *Then, on the event that the matrix $\mathbf{Z}$ has full rank K, we have, $\widehat{y}_x = \widehat{y}_z$ and $R(\widehat{\alpha}) = \mathbb{E}_{(X,y)}[(X^\top\widehat{\alpha} - y)^2] = \mathbb{E}_{(Z,y)}[(Z^\top\hat{\beta} - y)^2]$.*

2. *Suppose that Assumption 3 also holds and that $n > C \cdot K$ for some large enough absolute constant $C > 0$. Then, with probability at least $1 - c/n$ for some absolute constant $c > 0$, $\mathbf{Z}$ has full rank K and*

$$R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim \sigma_\varepsilon^2 \frac{K \log n}{n} \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\varepsilon}}[R(\widehat{\alpha})] - \sigma_\varepsilon^2 \gtrsim \sigma_\varepsilon^2 \frac{K}{n}. \tag{25}$$

The risk bounds (25) are the same as the standard risk bounds for prediction in linear regression in K dimensions with observable design, despite A not being known under model (5). We note that, since $\operatorname{rank}(\mathbf{X}) = K < n$, $\mathbf{y}$ may not lie in the range of $\mathbf{X}$ and so $\widehat{\alpha}$ may not interpolate. Nonetheless, under model (5), with $E \neq 0$ and in the interpolating regime, we expect that the prediction performance of $\widehat{y}_x$ will still approximately mimic that of $\widehat{y}_z$ as long as the signal, as measured by $\lambda_K(A^\top \Sigma_Z A)$, is strong relative to the noise, as measured by $\|\Sigma_E\|$. The next section is devoted to the detailed study of this fact.

Finally, another explanation of the perhaps surprisingly good performance of the GLS is that it coincides with Principal Component Regression (PCR), see, e.g., Stock and Watson (2002a), in the case when $\Sigma_E = 0$. Indeed, this is a natural and practical prediction method when the covariance matrix Σ_X has an approximately low rank. If $\Sigma_E = 0$, then $\Sigma_X = A\Sigma_Z A^\top$ has rank of at most K and so is exactly low rank. In PCR, the response $\mathbf{y}$ is regressed onto the first K principal components of the data matrix $\mathbf{X}$ to estimate a vector of coefficients $(\mathbf{X}\widehat{U}_K)^+\mathbf{y}$. Here $\widehat{U}_K \in \mathbb{R}^{p \times K}$ has columns equal to the first K eigenvectors of the sample covariance matrix $\mathbf{X}^\top\mathbf{X}/n$. A new response y is then predicted by $\widehat{\alpha}_{\mathrm{PCR}}^\top X$, where $\widehat{\alpha}_{\mathrm{PCR}} \coloneqq \widehat{U}_K(\mathbf{X}\widehat{U}_K)^+\mathbf{y}$ and X is the new feature vector. The following lemma states that the PCR and GLS predictors coincide when $\Sigma_E = 0$.

Lemma 13. *Define $\widehat{\alpha}_{\mathrm{PCR}} \coloneqq \widehat{U}_K(\mathbf{X}\widehat{U}_K)^+\mathbf{y}$. On the event $\{\operatorname{rank}(\mathbf{X}) = K\}$, $\widehat{\alpha} = \widehat{\alpha}_{\mathrm{PCR}}$. In particular, when $\Sigma_E = 0$, $K > C \cdot n$, and Assumptions 2 & 3 hold, $\widehat{\alpha} = \widehat{\alpha}_{\mathrm{PCR}}$ with probability at least $1 - c/n$ for some absolute constant $c > 0$.*

Thus, the prediction $\widehat{\alpha}_{\mathrm{PCR}}^\top X$ of y based on PCR is exactly equal to the prediction $\widehat{\alpha}^\top X$ based on the GLS, in the case when $\Sigma_E = 0$. Given that PCR is a natural and widely used prediction method in this setting, this further explains the performance of the GLS, at least when $\Sigma_E = 0$.

4.2 Approximate Adaptation of Interpolating Predictors in Factor Regression

In this section we present our main results on the excess risk of prediction with $\widehat{\alpha}$, relative to the two benchmarks in Section 3.2 above, under the factor regression model (5) with $E \neq 0$.

Our main result, Theorem 16 below, shows that despite the fact that $\widehat{\alpha}$ interpolates, in that $\mathbf{X}\widehat{\alpha} = \mathbf{y}$ (Proposition 14), and that $\|\widehat{\alpha}\| \to 0$ (Lemma 15), the excess risks can vanish as a result of approximate adaptation to the embedded low-dimensional structure of (5). The estimator $\widehat{\alpha}$ is guaranteed to interpolate the data whenever $\operatorname{rank}(\mathbf{X}) = n$, or equivalently, the smallest singular value $\sigma_n(\mathbf{X}) > 0$. The next proposition shows that the following set of conditions in terms of n, K and $\mathrm{r_e}(\Sigma_E)$ guarantee this. Proofs for this section are contained in Appendix C.2.

Proposition 14. *Under model (5), suppose that Assumptions 2 and 3 hold, and that* $\mathrm{r_e}(\Sigma_E) > C \cdot n$ *for some* $C > 0$ *large enough. Then, with probability at least* $1 - c/n$*, for some* $c > 0$*,*

$$\sigma_n^2(\mathbf{X}) \gtrsim \mathrm{tr}(\Sigma_E) > 0,$$

and thus, in particular, $\widehat{\alpha}$ *interpolates:* $\mathbf{X}\widehat{\alpha} = \mathbf{y}$*.*

General existing bounds of the type $\sigma_n(\mathbf{X}) \gtrsim (\sqrt{p} - \sqrt{n})$ are by now well established in random matrix theory (Rudelson and Vershynin, 2009). When $p > C \cdot n$ for some $C > 1$ and the entries of $\mathbf{X}$ are i.i.d. sub-Gaussian with zero mean and unit variance, Theorem 1.1 in Rudelson and Vershynin (2009) implies that $\sigma_n^2(\mathbf{X}) \gtrsim p$ with high probability. By comparison, Proposition 14 holds for $\mathbf{X}$ with i.i.d. sub-Gaussian rows with covariance matrix $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$.

The following result shows that as in the noiseless case $\Sigma_E = 0$ of Lemma 11, $\|\widehat{\alpha}\| \to 0$, mimicking the behavior of the best linear predictor α^*. We proved in Lemma 7 and Corollary 9 that $\|\alpha^*\| \to 0$ when $\lambda_K(A\Sigma_Z A^\top)$ grows faster than $\|\beta\|_{\Sigma_Z}^2$ as $p \to \infty$; we will need here the additional assumption that $n \log n / \mathrm{r_e}(\Sigma_E) \to 0$ to guarantee $\|\widehat{\alpha}\| \to 0$ as $n \to \infty$. The proof uses Proposition 14, which requires that the effective rank $\mathrm{r_e}(\Sigma_E)$ is larger than a constant times n.

Lemma 15. *Under model (5), suppose that Assumptions 2 and 3 hold and* $n > C \cdot K$ *and* $\mathrm{r_e}(\Sigma_E) > C \cdot n$ *hold, for some* $C > 0$*. Then, with probability exceeding* $1 - c/n$*, for some* $c > 0$*,*

$$\|\widehat{\alpha}\|^2 \lesssim \frac{1}{\lambda_K(A\Sigma_Z A^\top)} \|\beta\|_{\Sigma_Z}^2 + \sigma_\varepsilon^2 \frac{n \log n}{\mathrm{r_e}(\Sigma_E)}. \tag{26}$$

Despite the fact that $\|\widehat{\alpha}\| \to 0$ under the conditions stated, we now show that $\widehat{\alpha}$ can outperform the null predictor $\mathbf{0}$. If $\lambda_K(A\Sigma_Z A^\top)$ grows faster than $\mathrm{tr}(\Sigma_E)/n$ and $K/n \to 0$, then Lemma 2 states that $\mathrm{r_e}(\Sigma_X)/n$ remains bounded, and Theorem 1 allows for the possibility that $\widehat{\alpha}$ has asymptotically lower risk than $\mathbf{0}$. Theorem 12 above showed that $R(\widehat{\alpha}) - \sigma_\varepsilon^2$ can in fact approach 0 under certain conditions when $E = 0$. The following result demonstrates that this can continue to hold even when $E \neq 0$.

Theorem 16 (Main result: Risk bound for factor regression)**.** *Under model (5), suppose that Assumptions 2 and 3 hold and* $n > C \cdot K$ *and* $\mathrm{r_e}(\Sigma_E) > C \cdot n$ *hold, for some* $C > 0$*. Then, with probability exceeding* $1 - c/n$*, for some* $c > 0$*,*

$$R(\widehat{\alpha}) - R(\alpha^*) \;\leq\; R(\widehat{\alpha}) - \sigma_\varepsilon^2 \;\lesssim\; \frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \cdot \frac{\mathrm{r_e}(\Sigma_E)}{n} + \sigma_\varepsilon^2 \frac{n \log n}{\mathrm{r_e}(\Sigma_E)} + \sigma_\varepsilon^2 \frac{K \log n}{n}. \tag{27}$$

Recall $\xi \coloneqq \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\|$ *is the signal-to-noise ratio.*

Remark 17. *Suppose* $n \gg \sigma_\varepsilon^2 K \log n$ *and* $\mathrm{r_e}(\Sigma_E) \gg \sigma_\varepsilon^2 n \log n$*. We then find that* $\widehat{\alpha}$ *interpolates by Proposition 14, and the behavior of* $\widehat{\alpha}$ *is determined by the eigenvalue* $\lambda_K(A\Sigma_Z A^\top)$ *or, equivalently, the signal-to-noise ratio* $\xi = \lambda_K(A\Sigma_Z A^\top))/\|\Sigma_E\|$*.*

(a) If $\lambda_K(A\Sigma_Z A^\top) \gg \mathrm{tr}(\Sigma_E)/n$*, then Lemma 2 implies that* $R(\widehat{\alpha})$ *need no longer approach the trivial null risk* $R(\mathbf{0})$*.*

(b) If $\lambda_K(A\Sigma_Z A^\top) \gg \|\beta\|^2_{\Sigma_Z}$, then Lemma 15 implies $\|\widehat{\alpha}\| \to 0$.

(c) If $\lambda_K(A\Sigma_Z A^\top) \gg \|\beta\|^2_{\Sigma_Z}\mathrm{tr}(\Sigma_E)/n$, then $R(\widehat{\alpha}) - \sigma_\varepsilon^2 \to 0$. Indeed, this assumption, together with $n \gg \sigma_\varepsilon^2 K \log n$ and $\mathrm{r_e}(\Sigma_E) \gg \sigma_\varepsilon^2 n \log n$, ensures that the right-hand side of the inequality (27) in Theorem 16 is asymptotically negligible.

The first inequality in (27) is an immediate consequence of the first part of Lemma 5 above. We now discuss the three terms appearing in the upper bound (27) of Theorem 16. A comparison with the risk bound in Theorem 12 above, where the feature noise E is equal to zero, reveals that the term $\sigma_\varepsilon^2 K \log(n)/n$ in (27) is equal to the risk of the oracle predictor $\widehat{y}_z$ up to the multiplicative $\log n$ factor, and is small when $K \ll n$. The first two terms can be viewed as bias and variance components, respectively, that capture the impact of non-zero Σ_E. The first term (bias) is proportional to the effective rank $\mathrm{r_e}(\Sigma_E)$, while the second term (variance) is inversely proportional to $\mathrm{r_e}(\Sigma_E)$. As such, the variance term is implicitly regularized by the feature noise E, while for the bias to be small, we need the signal-to-noise ratio ξ to be sufficiently large. For example, suppose that the eigenvalues of Σ_Z and Σ_E are constant, that is, $c_1 \le \lambda_K(\Sigma_Z) \le \|\Sigma_Z\| \le C_1$ and $c_2 < \lambda_p(\Sigma_E) \le \|\Sigma_E\| < C_2$, for some $c_1, c_2, C_1, C_2 \in (0, \infty)$, both standard assumptions in factor models. Then,

$$\mathrm{r_e}(\Sigma_E) \asymp p, \quad \text{and} \quad \xi = \frac{\lambda_K(A\Sigma_Z A^\top)}{\|\Sigma_E\|} \gtrsim \lambda_K(A^\top A). \tag{28}$$

Provided β has uniformly bounded entries $|\beta_i| \le C$, $\|\beta\|^2_{\Sigma_Z} \le C_1 \cdot C^2 \cdot K$, and the bias term in (27) can be bounded as

$$B_Z \coloneqq \frac{\|\beta\|^2_{\Sigma_Z}}{\xi} \cdot \frac{\mathrm{r_e}(\Sigma_E)}{n} \lesssim \frac{Kp}{n \cdot \lambda_K(A^\top A)}; \tag{29}$$

it thus approaches zero whenever

$$\lambda_K(A^\top A) \gg \frac{Kp}{n}. \tag{30}$$

We mention that the examples of A in Remark 3 of Section 3.1 all imply (30), provided $K \ll n$ in cases 1 and 2 (since there $\lambda_K(A^\top A) \gtrsim p$), and $K^2 \ll n$ in case 3 (since there $\lambda_K(A^\top A) \gtrsim p/K$).

We summarize this discussion in Corollary 18 below.

Corollary 18. *Under the same conditions as in Theorem 16, suppose, in particular, that $\lambda_K(\Sigma_Z)$ and $\|\Sigma_E\|$ are constant, $\mathrm{r_e}(\Sigma_E) \asymp p$, and $\|\beta\|^2_{\Sigma_Z} \lesssim K$. Then, with probability at least $1 - c/n$, for some absolute constant $c > 0$,*

$$R(\widehat{\alpha}) - R(\alpha^*) \le R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim \frac{K}{\lambda_K(A^\top A)} \times \frac{p}{n} + \sigma_\varepsilon^2 \left(\frac{n}{p} + \frac{K}{n}\right) \log n. \tag{31}$$

In particular, if $\lambda_K(A^\top A) \gtrsim p/K$, and with probability at least $1 - c/n$, for some absolute constant $c > 0$,

$$R(\widehat{\alpha}) - R(\alpha^*) \le R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim \frac{K^2}{n} + \sigma_\varepsilon^2 \left(\frac{n}{p} + \frac{K}{n}\right) \log n. \tag{32}$$

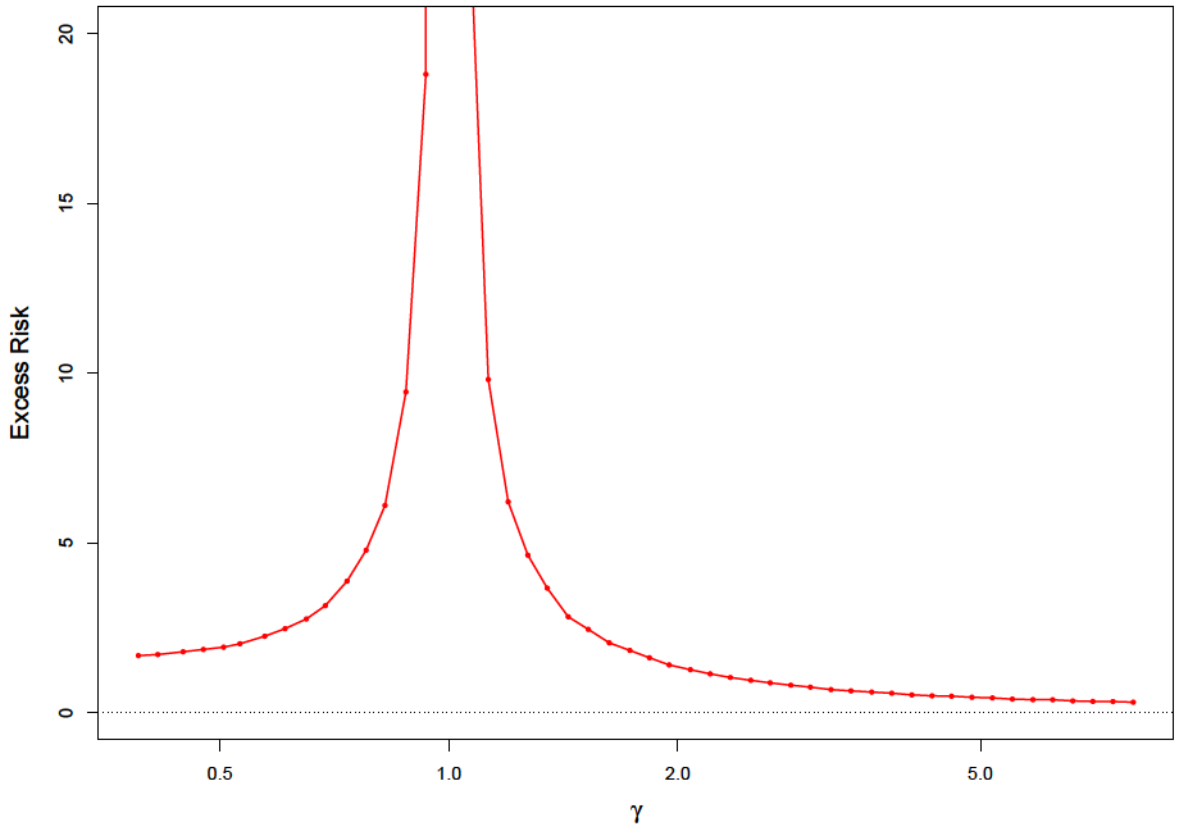

Figure 1: Excess prediction risk $R(\widehat{\alpha}) - \sigma_\varepsilon^2$ of the minimum-norm predictor under the factor regression model as a function of $\gamma = p/n$. Here K increases linearly from 16 to 64, $n = \lfloor K^{1.5} \rfloor$ and thus increases from 64 to 512, and p increases from 33 to 4066. Further, $\Sigma_E = I_p$, $\Sigma_Z = I_K$, $\beta = (1, \ldots, 1)^\top$, and $A = \sqrt{p} \cdot V_K$, where V_K is generated by taking the first K rows of a randomly generated $p \times p$ orthogonal matrix V.

Figure 1 illustrates the risk behavior proved in Theorem 16. Note the descent towards zero in the regime $\gamma := p/n > 1$. For completeness, we also provide a bound on the risk $R(\widehat{\alpha})$ for the low-dimensional case $p \ll n$, under model (5), in Appendix D.3.

4.3 Comparison to Existing Results

The recent paper Bartlett et al. (2020) gives a bias-variance type bound on the excess prediction risk of the minimum-norm predictor $\widehat{y}_x = X^\top \widehat{\alpha}$ considered in this work. In contrast to our study, Bartlett et al. (2020) does not consider model (5), and in fact assumes $\mathbb{E}[y|X] = X^\top \theta$ for some $\theta \in \mathbb{R}^p$, which is typically not satisfied under (5) when (X, y) are sub-Gaussian, but not Gaussian.

Regime	Bias in Theorem 16	Bias in Theorem 4 of Bartlett et al. (2020)	Common variance
$p \geq n \cdot \xi$	$\|\beta\|^2_{\Sigma_Z} \cdot p/(n \cdot \xi)$	$\|\beta\|^2_{\Sigma_Z} \cdot p/(n \cdot \xi)$	$\sigma_\varepsilon^2 \log n \{(n/p) + (K/n)\}$
$p \ll n \cdot \xi$	$\|\beta\|^2_{\Sigma_Z} \cdot p/(n \cdot \xi)$	$\|\beta\|^2_{\Sigma_Z} \cdot \sqrt{p/(n \cdot \xi)}$	
$\xi \approx p$, $\|\beta\|^2_{\Sigma_Z} \approx K$	K/n	$K/\sqrt{n}$	
$\xi \approx p$, $\|\beta\|^2_{\Sigma_Z} \approx K$, $K \approx n^{3/4}$	$n^{-1/4}$	$n^{1/4}$	

Table 2: Comparison of risk bounds for Gaussian data.

When the data are jointly Gaussian this assumption is, however, satisfied under model (5). For this common case, Table 2 compares the respective bounds on the bias and variance terms corresponding to our Theorem 16 and Theorem 4 of Bartlett et al. (2020), respectively. Again, we emphasize that the results from Bartlett et al. (2020) do not hold in general for our modeling setup, but can be used to obtain the bounds in Table 2 in the Gaussian case. The entries in the second column of Table 2 correspond to the bias in Bartlett et al. (2020) under model (5), simplified in this table for ease of comparison.[1]

In the setting of this comparison, the variance terms in our Theorem 16 and the bound in Bartlett et al. (2020) have the same rate, which we display in the third column of Table 2. From the first row of Table 2 we see that when $p \geq n \cdot \xi$, the bias terms match as well. However, this is not an interesting regime, as $p \ll n \cdot \xi$ is a necessary condition for either bound to converge to zero (assuming $\|\beta\|^2_{\Sigma_Z}$ is bounded below). In this case, the second row of Table 2 shows that the bias in Bartlett et al. (2020) becomes $\|\beta\|^2_{\Sigma_Z}\sqrt{p/(n \cdot \xi)}$, which is larger than our bias bound in Theorem 16 by a factor of $\sqrt{n \cdot \xi/p}$. From the second row we see that indeed, the upper bound on the excess risk in Bartlett et al. (2020) can diverge while our bound in Theorem 16 vanishes. For instance, if β is a non-sparse vector in $\mathbb{R}^K$ with $\|\beta\|^2_{\Sigma_Z} \approx K$, this phenomenon occurs if the signal-to-noise ratio ξ lies in the range $Kp/n \lesssim \xi \lesssim K^2 p/n$. This illustrates that the general bound provided in Bartlett et al. (2020) is not always tight.

The third row of Table 2 compares the bias rates in the simplified case when $\|\beta\|^2_{\Sigma_Z} \approx K$ and $\xi \approx p$. The fourth row gives the rates under the further assumption that $K \approx n^{3/4}$, a concrete example of when our rate converges and that of Bartlett et al. (2020) diverges. Further details and discussion on the comparison of these two results are deferred to Appendix C.4.

A latent factor regression model similar to (5) has also been studied in Section 7 of Mei and Montanari (2019) for the ridge regression estimator that minimizes the fit $\|\mathbf{y} -$

1. For simplicity, we assume for this comparison that the matrices Σ_X and Σ_E are invertible and that the condition numbers $\kappa(\Sigma_E)$ and $\kappa(A\Sigma_Z A^\top)$ are bounded above by an absolute constant. Consequently, the effective rank $\mathrm{r_e}(\Sigma_E)$ satisfies $c \cdot p \leq \mathrm{r_e}(\Sigma_E) \leq p$, for some $c \in (0, 1)$.

$\mathbf{X}a\|^2 + \lambda\|a\|^2$ for any $\lambda > 0$ (strict). Their model is a particular case of our model (5), with $\Sigma_E = \sigma_E^2 I_p$, $\Sigma_Z = \sigma_Z^2 I_K$, up to an offset on X so that in their case, $|\mathbb{E}[X]| > 0$. Clearly, our estimator $\widehat{\alpha}$ can be viewed as the limiting case $\lambda = 0$ of ridge regression. Our results are difficult to compare directly since the analysis in Mei and Montanari (2019) is asymptotic with $p/K \to \psi_1$ and $n/K \to \psi_2$ for two absolute constants $\psi_1, \psi_2 \in (0, \infty)$. Nevertheless, Theorem 7 and Figure 9 of Mei and Montanari (2019) also show that the excess risk $R(\widehat{\alpha}) - \sigma_\varepsilon^2$ is small in the large ψ_1/ψ_2 (corresponding to a large p/n) regime, in line with our assessment.

4.4 Comparison to Other Predictors

In Lemma 13 of Section 4.1 above we showed that in the case of noiseless features, when $\Sigma_E = 0$, the regression vector $\widehat{\alpha}_{\mathrm{PCR}}$ obtained by PCR is exactly equal to the GLS regression vector $\widehat{\alpha}$ on the event $\{\mathrm{rank}(\mathbf{Z}) = K\}$, which holds with probability at least $1 - c/n$ for some universal constant $c > 0$. In this section we show that when $\Sigma_E \neq 0$, the minimum-norm estimator $\widehat{\alpha}$ is competitive even with the stylized version $\widetilde{\alpha}_{\mathrm{PCR}} \coloneqq U_K(\mathbf{X}U_K)^+\mathbf{y}$ of PCR under the factor regression model setting (5) and in the high-dimensional regime $p \gg n$. This is a toy estimator as it uses the unknown dimension K and unknown matrix U_K, composed of the first K eigenvectors of the population covariance matrix Σ_X, in place of estimates $\widehat{K}$ and $\widehat{U}_{\widehat{K}}$, respectively. We provide a simple proof, found in Appendix C.3, of the following risk bound for $R(\widetilde{\alpha}_{\mathrm{PCR}})$. For a detailed comparison of PCR and the GLS, see Bing et al. (2020), which analyzes the PCR predictor with the empirical matrix $\widehat{U}_{\widehat{K}}$, for a new, data adaptive, estimator $\widehat{K}$ of K.

Theorem 19. *Under model (5), suppose that (X, y) are jointly Gaussian and that Assumption 2 holds. Then, if $n > C \cdot K \log n$ for some $C > 0$ large enough, with probability at least $1 - c/n$,*

$$R(\widetilde{\alpha}_{\mathrm{PCR}}) - \sigma_\varepsilon^2 \lesssim \|\Sigma_E\| \cdot \|\alpha^*\|^2 \frac{p}{n} + R(\alpha^*) \frac{K \log(n)}{n} \tag{33}$$

In particular, if $\Sigma_E = 0$, we obtain

$$R(\widetilde{\alpha}_{\mathrm{PCR}}) - \sigma_\varepsilon^2 \lesssim \sigma_\varepsilon^2 \frac{K \log(n)}{n} \tag{34}$$

while, if $\lambda_p(\Sigma_E) > 0$,

$$R(\widetilde{\alpha}_{\mathrm{PCR}}) - \sigma_\varepsilon^2 \lesssim \kappa(\Sigma_E) \frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \frac{p}{n} + \sigma_\varepsilon^2 \frac{K \log n}{n}, \tag{35}$$

where $\kappa(\Sigma_E) \coloneqq \lambda_1(\Sigma_E)/\lambda_p(\Sigma_E)$ is the condition number of the matrix Σ_E.

Provided $\kappa(\Sigma_E)$ is bounded above by an absolute constant, the upper bounds for the minimum-norm and PCR predictors are comparable. Indeed, when $\kappa(\Sigma_E) < C < \infty$, the risk bound of Theorem 16 for the GLS $\widehat{\alpha}$ takes the form

$$R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim \frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \frac{p}{n} + \sigma_\varepsilon^2 \log n \left(\frac{K}{n} + \frac{n}{p} \right). \tag{36}$$

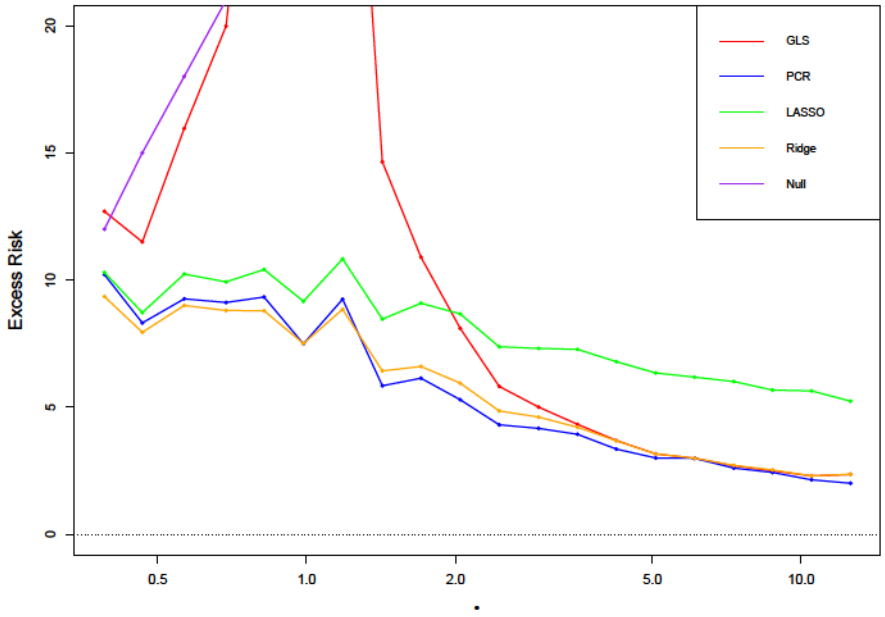

Figure 2: Excess prediction risk of GLS, PCR, LASSO, Ridge regression, and the null predictor as a function of $\gamma = p/n$. Here K increases linearly from 12 to 69, $n = \lfloor K^{1.5} \rfloor$ and thus increases from 41 to 573, and p increases from 16 to 7215. Further, $\Sigma_E = I_p$, $\Sigma_Z = I_K$, $\beta = (1, \ldots, 1)^\top$, and A is generated by sampling each entry iid from $N(0, 1/\sqrt{K})$.

The additional term $\sigma_\varepsilon^2 n \log n/p$ in this bound is absent in the PCR prediction bound (35) above, but in the regime $p \gg n$ it can become negligible. It is perhaps surprising that under the factor regression model, the interpolator $\widehat{\alpha}$ can not only provide consistent prediction, but can in fact have excess risk comparable to a genuine K-dimensional predictor widely used in practice and tailored to the problem setting. This is despite the fact that the GLS interpolates the data (when $\operatorname{rank}(\mathbf{X}) = n$) and requires no tuning parameters or knowledge of the underlying dimension K. We emphasize that we do not claim that the GLS is necessarily a superior predictor to PCR in this setting. Rather, we observe the perhaps surprising fact that these two methods are comparable under the conditions stated.

Figure 2 plots the excess prediction risk of the GLS and PCR predictors. We also include the excess prediction risks of the LASSO, Ridge regression, and the null estimator $\mathbf{0}$ in this figure for comparison. The tuning parameters for LASSO and Ridge regression were chosen by cross-validation. We see that the peak in the GLS risk at $\gamma = p/n = 1$ is not present in the PCR, LASSO and Ridge risks. This is due to the fact that these methods are regularized at this point, and in particular do not interpolate the training data. As γ increases, and thus $p \gg n$, the GLS risk approaches the PCR risk, as indicated by the discussion above. The plot shows how the Ridge risk also approaches the common value of the PCR and GLS risks. Recalling that GLS is a limiting case of Ridge regression with

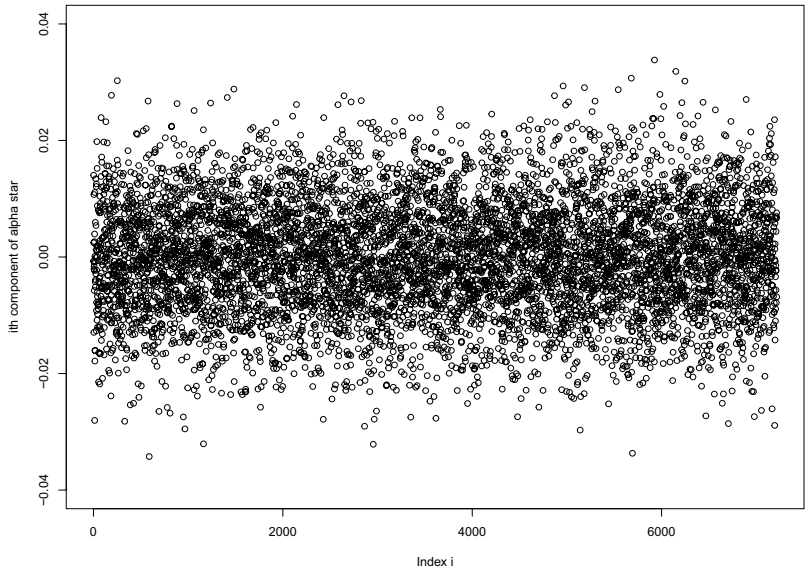

Figure 3: A scatter plot of the components of α^*, from the point in the simulation of Figure 2 with the largest value of γ. Here $p = 7215$, $K = 69$, $\Sigma_E = I_p$, $\Sigma_Z = I_K$, and A is generated by sampling each entry iid from $N(0, 1/\sqrt{K})$.

regularization parameter $\lambda \to 0$, this suggests that for $p \gg n$, in our setting, the optimal choice of regularization parameter for ridge regression approaches zero Mei and Montanari (2019); Hastie et al. (2019).

We plot the coefficients of α^* in Figure 3 for the case $p = 7215$ and $K = 69$. We can see that α^* is clearly non-sparse, which explains the inferior performance of the LASSO in this setting.

For completeness, we contrast the above simulation setting in which α^* is non-sparse with special case in which α^* is in fact K-sparse. In this case, we take the matrix A with columns equal to the canonical basis vectors $e_1, \ldots, e_K \in \mathbb{R}^p$, multiplied by $\sqrt{p}$, and we set $\beta = (1, \ldots, 1)^\top$, $\Sigma_Z = I_K$ and $\Sigma_E = I_p$. Then $A^\top A = pI_K$ and α^* is K-sparse since, by (18) of Remark 8,

$$\alpha_i^* = \begin{cases} \sqrt{p}/(p+1) & \text{for } i = 1, \ldots, K \\ 0 & \text{for } i = K+1, \ldots, p \end{cases}.$$

Figure 4 plots the excess risk of the GLS and other predictors for these model settings. We see that in this sparse setting the LASSO performs well, as expected, with its excess risk approximately equal to that of PCR for $p \gg n$, both of which do slightly better than GLS and Ridge. While LASSO and PCR outperform GLS in this case, we note that the excess risk of the GLS still decreases towards zero, and performs perhaps surprisingly well relative to the LASSO, given that the LASSO is specifically tailored to this exactly sparse setting. Moreover, we emphasize that for more generic choices of model parameters, α^* will

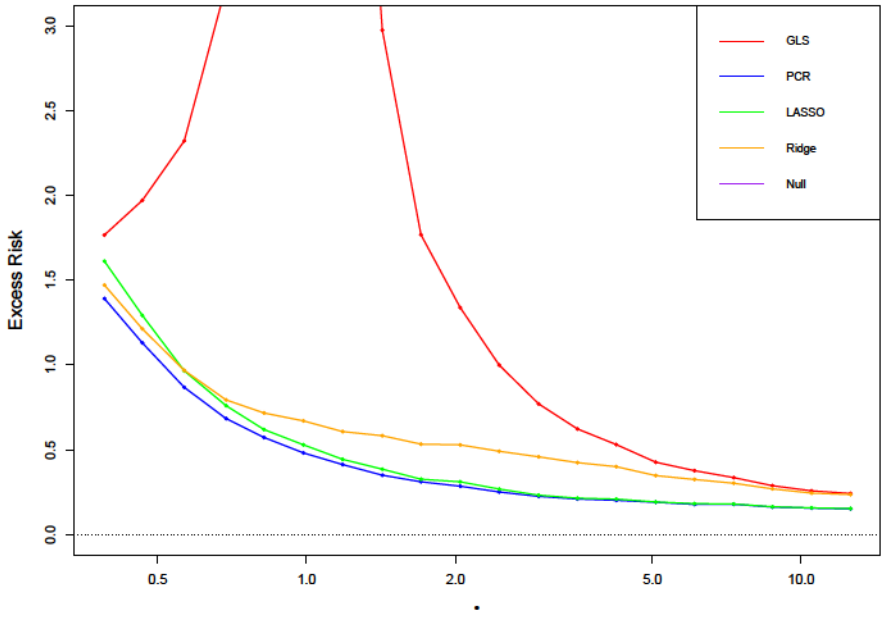

Figure 4: Excess prediction risk of GLS, PCR, LASSO, Ridge regression, and the null predictor as a function of $\gamma = p/n$. Null risk is not visible on plot since it is larger than the maximum plotted value. Here K increases linearly from 12 to 69, $n = \lfloor K^{1.5} \rfloor$ and thus increases from 41 to 573, and p increases from 16 to 7215. Further, $\Sigma_E = I_p$, $\Sigma_Z = I_K$, $\beta = (1, \ldots, 1)^\top$, and A has columns equal to the canonical basis vectors $e_1, \ldots, e_K \in \mathbb{R}^p$, multiplied by $\sqrt{p}$.

not necessarily be sparse or even approximately sparse, and we should expect the GLS to outperform the LASSO (see Remark 8 for further comment).

The take-home message is that for $\gamma = p/n$ large enough, the GLS is a surprisingly competitive predictor, given its interpolating property, and in fact performs as well in the generic setting of Figure 2 as the PCR predictor chosen with the unknown, optimal number of components K, in addition to Ridge regression with tuning parameter chosen by cross-validation. Even when the model parameters are carefully chosen so that the best linear predictor α^* is K-sparse, the GLS performs not much worse than the LASSO, which is tailored to this setting, provided that p is very large.

Acknowledgments

We thank the anonymous referees for their many insightful and helpful suggestions. During the completion of this work, Bunea and Wegkamp were supported in part by NSF grant DMS 2015195, and Seth Strimas-Mackey was partially supported by NSERC PGS-D.

Appendix A. Proofs for Section 2

A.1 Proof of Theorem 1

We work on the event

$$\mathcal{K} \coloneqq \left\{\sigma_n^2(\mathbf{X}) \gtrsim \mathrm{tr}(\Sigma_X),\ \|\mathbf{y}\|^2 \lesssim n\sigma_y^2\right\}. \tag{37}$$

On this event, recalling $\widehat{\alpha} = \mathbf{X}^+\mathbf{y}$ and invoking identity (137) in Appendix E,

$$\|\widehat{\alpha}\|^2 \le \|\mathbf{X}^+\|^2\|\mathbf{y}\|^2 = \frac{\|\mathbf{y}\|^2}{\sigma_n^2(\mathbf{X})} \lesssim \sigma_y^2\frac{n}{\mathrm{tr}(\Sigma_X)}. \tag{38}$$

By Lemma 20 below,

$$\left|\frac{R(\theta)}{R(\mathbf{0})} - 1\right| \le \frac{\|\theta\|_{\Sigma_X}^2}{R(\mathbf{0})} + 2\sqrt{\frac{\|\theta\|_{\Sigma_X}^2}{R(\mathbf{0})}} \le \|\Sigma_X\|\frac{\|\theta\|^2}{R(\mathbf{0})} + 2\sqrt{\|\Sigma_X\|\frac{\|\theta\|^2}{R(\mathbf{0})}}$$

for any vector $\theta \in \mathbb{R}^p$. Combining this with (38) and recalling that $\sigma_y^2 = \mathbb{E}[y^2] = R(\mathbf{0})$, we find that on $\mathcal{K}$,

$$\left|\frac{R(\widehat{\alpha})}{R(\mathbf{0})} - 1\right| \lesssim \frac{n}{\mathrm{r_e}(\Sigma_X)} + \sqrt{\frac{n}{\mathrm{r_e}(\Sigma_X)}}$$

Setting $C' = \max(C, 1)$, when $\mathrm{r_e}(\Sigma_X) > C'n \ge n$, so $n/\mathrm{r_e}(\Sigma_X) > 1$, we find

$$\frac{n}{\mathrm{r_e}(\Sigma_X)} + \sqrt{\frac{n}{\mathrm{r_e}(\Sigma_X)}} \le 2\sqrt{\frac{n}{\mathrm{r_e}(\Sigma_X)}}.$$

Thus, on $\mathcal{K}$,

$$\left|\frac{R(\widehat{\alpha})}{R(\mathbf{0})} - 1\right| \lesssim \sqrt{\frac{n}{\mathrm{r_e}(\Sigma_X)}}.$$

All that remains is to bound the probability of $\mathcal{K}$. To this end, note that since we suppose Assumption 1 holds, we have $\mathbf{X} = \tilde{\mathbf{X}}\Sigma_X^{1/2}$, and thus

$$\sigma_n^2(\mathbf{X}) = \lambda_n(\mathbf{X}\mathbf{X}^\top) = \lambda_n(\tilde{\mathbf{X}}\Sigma_X\tilde{\mathbf{X}}),$$

where $\tilde{\mathbf{X}}$ has i.i.d. entries that have zero mean, unit variance, and sub-Gaussian constants bounded by an absolute constant. Theorem 21 below thus implies that if $\mathrm{r_e}(\Sigma_X) > C \cdot n$ for $C > 0$ large enough, then with probability at least $1 - 2e^{-cn}$,

$$\sigma_n^2(\mathbf{X}) \ge \mathrm{tr}(\Sigma_X)/2 - c_0\|\Sigma_X\|n = \mathrm{tr}(\Sigma_X) \cdot [1/2 - c_0 n/\mathrm{r_e}(\Sigma_X)].$$

Using that $n/\mathrm{r_e}(\Sigma_X) < 1/C$ and choosing C large enough,

$$\mathbb{P}(\sigma_n^2(\mathbf{X}) \gtrsim \mathrm{tr}(\Sigma_X)) \ge 1 - 2e^{-cn}. \tag{39}$$

By Assumption 1, $\mathbf{y} = \sigma_y\tilde{\mathbf{y}}$. Since $\tilde{y}_1, \ldots, \tilde{y}_n$ have zero mean and sub-Gaussian constants bounded by an absolute constant, Bernstein's inequality (Corollary 2.8.3 of Vershynin (2019)) implies that

$$\mathbb{P}(\|\tilde{\mathbf{y}}\|^2 \gtrsim n) = \mathbb{P}\left(\left|\sum_{i=1}^n \tilde{y}_i^2\right| \gtrsim n\right) \le 2e^{-2cn}.$$

Thus,

$$\mathbb{P}(\|\mathbf{y}\|^2 \gtrsim \sigma_y^2 n) = \mathbb{P}(\sigma_y^2\|\tilde{\mathbf{y}}\|^2 \gtrsim \sigma_y^2 n) = \mathbb{P}(\|\tilde{\mathbf{y}}\|^2 \gtrsim n) \le 2e^{-2cn}.$$

Combining this with (39) establishes that $\mathbb{P}(\mathcal{K}) \ge 1 - ce^{-c'n}$, thus completing the proof. ■

A.2 Lemma 20 and Theorem 21

The proof of Theorem 1 above made crucial use of the following lemma and theorem.

Lemma 20. *For any vector $\theta \in \mathbb{R}^p$,*

$$\left|\frac{R(\theta)}{R(\mathbf{0})} - 1\right| \leq \frac{\|\theta\|_{\Sigma_X}^2}{R(\mathbf{0})} + 2\sqrt{\frac{\|\theta\|_{\Sigma_X}^2}{R(\mathbf{0})}}. \tag{40}$$

Proof We first show that $\Sigma_X \alpha^* = \Sigma_{Xy}$, where $\Sigma_{Xy} \coloneqq \mathbb{E}[Xy]$ and $\alpha^* \coloneqq \Sigma_X^+ \Sigma_{Xy}$. To this end, observe that

$$\begin{aligned}
\mathrm{Cov}((I - \Sigma_X \Sigma_X^+)X) &= (I_p - \Sigma_X \Sigma_X^+)\mathbb{E}[XX^\top](I_p - \Sigma_X \Sigma_X^+) \\
&= (I_p - \Sigma_X \Sigma_X^+)\Sigma_X(I_p - \Sigma_X^+ \Sigma_X) \\
&= 0,
\end{aligned}$$

where we use that $\Sigma_X \Sigma_X^+ \Sigma_X = \Sigma_X$ (see Appendix E). Thus $(I_p - \Sigma_X \Sigma_X^+)X = 0$ a.s., so

$$\Sigma_X \alpha^* = \Sigma_X \Sigma_X^+ \Sigma_{Xy} = \mathbb{E}[\Sigma_X \Sigma_X^+ Xy] = \mathbb{E}[Xy] = \Sigma_{Xy}. \tag{41}$$

Fixing $\theta \in \mathbb{R}^p$, we have

$$\begin{aligned}
R(\theta) - R(\mathbf{0}) &= \mathbb{E}[(X^\top \theta - y)^2] - \mathbb{E}[y^2] \\
&= \theta^\top \mathbb{E}[XX^\top]\theta - 2\theta^\top \mathbb{E}[Xy] \\
&= \|\theta\|_{\Sigma_X}^2 - 2\theta^\top \Sigma_{Xy} \\
&= \|\theta\|_{\Sigma_X}^2 - 2\theta^\top \Sigma_X \alpha^* \qquad \text{(by (41))},
\end{aligned}$$

so by the Cauchy-Schwarz inequality,

$$|R(\theta) - R(\mathbf{0})| \leq \|\theta\|_{\Sigma_X}^2 + 2\|\theta\|_{\Sigma_X}\|\alpha^*\|_{\Sigma_X}. \tag{42}$$

Next observe that

$$R(\mathbf{0}) = \mathbb{E}[y^2] = \mathbb{E}(y - X^\top \alpha^* + X^\top \alpha^*)^2 = R(\alpha^*) + \|\alpha^*\|_{\Sigma_X}^2 \geq \|\alpha^*\|_{\Sigma_X}^2,$$

where we use that by (41),

$$\mathbb{E}(X^\top \alpha^*)(X^\top \alpha^* - y) = \alpha^{*\top} \Sigma_X \alpha^* - \alpha^{*\top} \Sigma_{Xy} = 0.$$

Thus, $\|\alpha^*\|_{\Sigma_X}^2 \leq R(\mathbf{0})$, so by (42),

$$|R(\theta) - R(\mathbf{0})| \leq \|\theta\|_{\Sigma_X}^2 + 2\|\theta\|_{\Sigma_X}\sqrt{R(\mathbf{0})}. \tag{43}$$

Dividing both sides by $R(\mathbf{0})$ gives the final result. ■

Theorem 21. *Suppose* $\mathbf{W}$ *is an* $n \times r$ *random matrix with independent subgaussian entries that have zero mean and unit variance. Then for any positive semi-definite matrix* $\Sigma \in \mathbb{R}^{r\times r}$ *and some* $c' > 0$ *large enough, with probability at least* $1 - 2e^{-cn}$,

$$\mathrm{tr}(\Sigma)/2 - c'(M^2+M^4)\|\Sigma\|n \le \lambda_n(\mathbf{W}\Sigma\mathbf{W}^\top) \le \lambda_1(\mathbf{W}\Sigma\mathbf{W}^\top) \le 3\mathrm{tr}(\Sigma)/2 + c'(M^2+M^4)\|\Sigma\|n,$$

where $M \coloneqq \max_{i,j} \|\mathbf{W}_{ij}\|_{\psi_2}$.[2]

A similar result for diagonal Σ has been derived in Lemma 9 of Bartlett et al. (2020). We make use of the Hanson-Wright inequality in our proof to deal with non-diagonal Σ. Theorem 4.6.1 in Vershynin (2019) provides similar two-sided bounds for the smallest and largest eigenvalue of $\mathbf{W}\Sigma\mathbf{W}^\top$, when $\Sigma = I_r$.

Proof We will prove that for some $c' \ge 1$,

$$\|\mathbf{W}\Sigma\mathbf{W}^\top - \mathrm{tr}(\Sigma)I_n\| \le c'(M^2+M^4)\|\Sigma\|n + \mathrm{tr}(\Sigma)/2 \tag{44}$$

with probability at least $1-2e^{-cn}$. Equation (44) implies that for any $v \in \mathbb{R}^n$ with $\|v\| = 1$,

$$|v^\top\mathbf{W}\Sigma\mathbf{W}^\top v - \mathrm{tr}(\Sigma)| \le c'(M^2+M^4)\|\Sigma\|n + \mathrm{tr}(\Sigma)/2,$$

and so

$$\mathrm{tr}(\Sigma)/2 - c'(M^2+M^4)\|\Sigma\|n \le v^\top\mathbf{W}\Sigma\mathbf{W}^\top v \le 3\mathrm{tr}(\Sigma)/2 + c'(M^2+M^4)\|\Sigma\|n.$$

Taking the minimum and maximum over $v \in S^{n-1}$ then gives the desired result.

We now prove (44). Let $\mathcal{N}$ be a 1/4-net of S^{n-1} with $|\mathcal{N}| \le 9^n$, which exists by Corollary 4.2.13 of Vershynin (2019). Then by Exercise 4.4.3 of Vershynin (2019),

$$\|\mathbf{W}\Sigma\mathbf{W}^\top - \mathrm{tr}(\Sigma)I_n\| = \sup_{v\in S^{n-1}} |v^\top\mathbf{W}\Sigma\mathbf{W}^\top v - \mathrm{tr}(\Sigma)| \le 2\sup_{v\in\mathcal{N}} |v^\top\mathbf{W}\Sigma\mathbf{W}^\top v - \mathrm{tr}(\Sigma)|, \tag{45}$$

where we use that $\mathbf{W}\Sigma\mathbf{W}^\top - \mathrm{tr}(\Sigma)I_n$ is symmetric in the first step.

Now fix $v \in S^{n-1}$ and define $B = \mathbf{W}^\top v \in \mathbb{R}^r$. Observe that B has mean zero entries that are independent because the columns of $\mathbf{W}$ are independent. Furthermore, by Proposition 2.6.1 of Vershynin (2019),

$$\|B_i\|_{\psi_2}^2 = \|\sum_j \mathbf{W}_{ji}v_j\|_{\psi_2}^2 \le C\sum_j \|\mathbf{W}_{ji}\|_{\psi_2}^2 v_j^2 \le \max_{li}\|\mathbf{W}_{li}\|_{\psi_2}^2 \sum_j v_j^2 = CM^2,$$

where we used $\|v\|^2 = 1$ in the last step. Thus, by the Hanson-Wright inequality (Theorem 6.2.1 in Vershynin (2019)),

$$\mathbb{P}\left(|B^\top\Sigma B - \mathbb{E}B^\top\Sigma B| \ge c_1M^2t\right) \le 2\exp\left\{-c_2\min\left(t/\|\Sigma\|, t^2/\|\Sigma\|_F^2\right)\right\}, \tag{46}$$

where we can choose $c_1 > 0$ large enough such that $c_2 \ge 12$.

2. We define the sub-Gaussian norm of any real-valued random variable U by $\|U\|_{\psi_2} \coloneqq \inf\{t > 0 : \mathbb{E}\exp(U^2/t) < 2\}$. We say U is sub-Gaussian when $\|U\|_{\psi_2} < \infty$.

Note that

$$\mathbb{E}B^\top \Sigma B = \sum_{i,j,k,l} \mathbb{E} v_i \mathbf{W}_{ij} \Sigma_{jl} \mathbf{W}_{kl} v_k = \sum_{ij} v_i^2 \Sigma_{jj} \mathbb{E}\mathbf{W}_{ij}^2 = \|v\|^2 \mathrm{tr}(\Sigma) = \mathrm{tr}(\Sigma), \tag{47}$$

where in the second step we use that $\mathbf{W}$ has independent mean zero entries, in the third step we use that $\mathbb{E}\mathbf{W}_{ij}^2 = 1$ for all i, j, and in the final step we use that $\|v\| = 1$.

Choosing $t = \|\Sigma\| n/2 + \sqrt{n\|\Sigma\|_F^2}/2$ in (46) and using that $c_2 \geq 12$, we observe that

$$c_2 t/\|\Sigma\| = c_2 n/2 + c_2\sqrt{n\|\Sigma\|_F^2}/(2\|\Sigma\|) \geq c_2 n/2 \geq 3n,$$

and

$$c_2 t^2/\|\Sigma\|_F^2 = c_2\left[n\|\Sigma\|/(2\|\Sigma\|_F) + \sqrt{n}/2\right]^2 \geq c_2 n/4 \geq 3n.$$

Thus,

$$\mathbb{P}\left(|B^\top \Sigma B - \mathrm{tr}(\Sigma)| \geq c_1 M^2 \|\Sigma\| n/2 + c_1 M^2 \sqrt{n\|\Sigma\|_F^2}/2\right) \leq 2e^{-3n}, \tag{48}$$

where we used (47). Finally, using

$$\|\Sigma\|_F^2 = \mathrm{tr}(\Sigma^2) \leq \|\Sigma\| \mathrm{tr}(\Sigma),$$

and the inequality $2ab \leq a^2 + b^2$,

$$c_1 M^2 \sqrt{n\|\Sigma\|_F^2}/2 \leq c_1 M^2 \sqrt{(c_1 M^2 n\|\Sigma\|)(\mathrm{tr}(\Sigma)/c_1 M^2)}/2 \leq c_1^2 M^4 n\|\Sigma\|/4 + \mathrm{tr}(\Sigma)/4.$$

Thus, by (48), and for $c' > 0$ large enough,

$$\mathbb{P}\left(|B^\top \Sigma B - \mathrm{tr}(\Sigma)| \geq c'(M^2 + M^4)\|\Sigma\| n + \mathrm{tr}(\Sigma)/4\right) \leq 2e^{-3n}. \tag{49}$$

Denoting $c'(M^2 + M^4)\|\Sigma\| n + \mathrm{tr}(\Sigma)/4$ by L, we thus have

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{W}\Sigma\mathbf{W}^\top - \mathrm{tr}(\Sigma) I_n\| \geq 2L\right) &\leq \mathbb{P}\left(2\sup_{v \in \mathcal{N}} |v^\top \mathbf{W}\Sigma\mathbf{W}^\top v - \mathrm{tr}(\Sigma)| \geq 2L\right) && \text{(by (45))} \\
&\leq \sum_{v \in \mathcal{N}} \mathbb{P}\left(|v^\top \mathbf{W}\Sigma\mathbf{W}^\top v - \mathrm{tr}(\Sigma)| \geq L\right) && \text{(union bound)} \\
&\leq 2 \times 9^n e^{-3n} && \text{(by (49))} \\
&= 2e^{n\log(9) - 3n} \leq 2e^{-cn},
\end{aligned}$$

where we define $c = 3 - \log(9) > 0$ in the last step. This shows (44) and completes the proof. ∎

Appendix B. Proofs for Section 3

B.1 Proof of Lemma 4 from Section 3.1

We will use $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$ and the min-max formula for eigenvalues,

$$\lambda_i(\Sigma_X) = \min_{S:\dim(S)=i} \max_{x\in S:\|x\|=1} x^\top \Sigma_X x, \tag{50}$$

where the minimum is taken over all linear subspaces $S \subset \mathbb{R}^p$ with dimension i. We prove the three points one by one.

1. Since for any $x \in \mathbb{R}^p$, $x^\top A\Sigma_Z A^\top x \geq 0$, we have

   $$x^\top \Sigma_X x \geq x^\top \Sigma_E x,$$

   so by (50), for any $i \in [p]$,

   $$\lambda_i(\Sigma_X) \geq \lambda_i(\Sigma_E) \geq \lambda_p(\Sigma_E) > c_2.$$

2. For any $x \in \mathbb{R}^p$,

   $$\begin{aligned} x^\top \Sigma_X x &= x^\top A\Sigma_Z A^\top x + x^\top \Sigma_E x \\ &\geq x^\top A\Sigma_Z A^\top x \\ &\geq \lambda_K(\Sigma_Z) x^\top AA^\top x \\ &\geq c_1 \cdot x^\top AA^\top x. \end{aligned}$$

   Plugging this into (50) with $i = K$, we find $\lambda_K(\Sigma_X) \geq c_1 \lambda_K(A^\top A)$ as claimed.

3. For any $x \in \mathbb{R}^p$, $x^\top \Sigma_E x \leq \|\Sigma_E\|$. Using this in (50), we find for any $i > K$,

   $$\lambda_i(\Sigma_X) \leq \|\Sigma_E\| + \lambda_i(A\Sigma_Z A^\top) = \|\Sigma_E\| < C_2,$$

   where in the second step we use that $\operatorname{rank}(A\Sigma_Z A^\top) \leq K$, so $\lambda_i(A\Sigma_Z A^\top) = 0$ for $i > K$. Combining this with $\lambda_i(\Sigma_X) > c_2$ from part 1 above completes the proof.

■

B.2 Proof of Lemma 5 from Section 3.2

Using $y = Z^\top \beta + \varepsilon$ and the fact that ε is independent of X and Z,

$$R(\alpha^*) = \mathbb{E}[(\alpha^{*\top} X - y)]^2 = \mathbb{E}[(\alpha^{*\top} X - Z^\top \beta)]^2 + \sigma_\varepsilon^2 \geq \sigma_\varepsilon^2,$$

which proves the first claim. Using $X = AZ + E$, we further find

$$R(\alpha^*) - \sigma_\varepsilon^2 = \mathbb{E}[(\alpha^{*\top} X - Z^\top \beta)]^2 = \alpha^{*\top} \Sigma_X \alpha^* + \beta^\top \Sigma_Z \beta - 2\alpha^{*\top} A\Sigma_Z \beta. \tag{51}$$

Now suppose Σ_E and Σ_Z are invertible as in the second claim. Then in particular,

$$\lambda_p(\Sigma_X) \geq \lambda_p(\Sigma_E) > 0,$$

so Σ_X is invertible and thus $\Sigma_X^+ = \Sigma_X^{-1}$. Also, $\Sigma_{Xy} = \mathbb{E}[Xy] = A\Sigma_Z\beta$, so

$$\alpha^* = \Sigma_X^+\Sigma_{Xy} = \Sigma_X^{-1}A\Sigma_Z\beta.$$

Defining $\bar{A} \coloneqq A\Sigma_Z^{1/2}$ and $\bar{\beta} \coloneqq \Sigma_Z^{1/2}\beta$, we have $\alpha^* = \Sigma_X^{-1}\bar{A}\bar{\beta}$. Plugging this into (51) and simplifying, we find

$$R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top \left[I_K - \bar{A}^\top\Sigma_X^{-1}\bar{A}\right]\bar{\beta}. \tag{52}$$

By the Woodbury matrix identity,

$$\Sigma_X^{-1} = (\bar{A}\bar{A}^\top + \Sigma_E)^{-1} = \Sigma_E^{-1} - \Sigma_E^{-1}\bar{A}(I_K + \bar{A}^\top\Sigma_E^{-1}\bar{A})^{-1}\bar{A}^\top\Sigma_E^{-1},$$

so letting $\bar{G} \coloneqq I_K + \bar{A}^\top\Sigma_E^{-1}\bar{A}$,

$$\bar{A}^\top\Sigma_X^{-1}\bar{A} = \bar{A}^\top\Sigma_E^{-1}\bar{A} - \bar{A}^\top\Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{A}^\top\Sigma_E^{-1}\bar{A}.$$

Now using $\bar{A}^\top\Sigma_E^{-1}\bar{A} = \bar{G} - I_K$, we find

$$\begin{aligned}\bar{A}^\top\Sigma_X^{-1}\bar{A} &= (\bar{G} - I_K) - (\bar{G} - I_K)\bar{G}^{-1}(\bar{G} - I_K)\\ &= \bar{G} - I_K - (I_K - \bar{G}^{-1})(\bar{G} - I_K)\\ &= \bar{G} - I_K - [\bar{G} - I_K - I_K + \bar{G}^{-1}]\\ &= I_K - \bar{G}^{-1}.\end{aligned}$$

Using this to simplify (52), we find

$$R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top\bar{G}^{-1}\bar{\beta} = \bar{\beta}^\top(I_K + \bar{A}\Sigma_E^{-1}\bar{A})^{-1}\bar{\beta}. \tag{53}$$

Letting $H \coloneqq \bar{A}\Sigma_E^{-1}\bar{A}$, we find

$$R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top H^{-1/2}(I_K + H^{-1})^{-1}H^{-1/2}\bar{\beta}. \tag{54}$$

For the lower bound, first observe that

$$R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top H^{-1/2}(I_K + H^{-1})^{-1}H^{-1/2}\bar{\beta} \geq \frac{\bar{\beta}^\top H^{-1}\bar{\beta}}{1 + \|H^{-1}\|} = \frac{\beta^\top(A\Sigma_E^{-1}A)^{-1}\beta}{1 + \lambda_K^{-1}(H)}.$$

Furthermore,

$$\lambda_K(H) = \lambda_K(\bar{A}^\top\Sigma_E^{-1}\bar{A}) \geq \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| = \xi, \tag{55}$$

so using this in the previous display,

$$R(\alpha^*) - \sigma_\varepsilon^2 \geq \frac{\beta^\top(A^\top\Sigma_E^{-1}A)^{-1}\beta}{1 + \xi^{-1}} = \frac{\xi}{1 + \xi}\cdot\beta^\top(A^\top\Sigma_E^{-1}A)^{-1}\beta.$$

To obtain the upper bound on $R(\alpha^*)$ we use

$$R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top H^{-1/2}(I_K + H^{-1})^{-1}H^{-1/2}\bar{\beta} \leq \frac{\bar{\beta}^\top H^{-1}\bar{\beta}}{1 + \lambda_K(H^{-1})} \leq \bar{\beta}^\top H^{-1}\bar{\beta} = \beta^\top(A\Sigma_E^{-1}A)^{-1}\beta,$$

where in the last step we use $\Sigma_Z^{1/2}H^{-1}\Sigma_Z^{1/2} = (A\Sigma_E^{-1}A)^{-1}$. Finally,

$$\beta^\top(A\Sigma_E^{-1}A)^{-1}\beta = \bar{\beta}^\top H^{-1}\bar{\beta} \leq \|\beta\|_{\Sigma_Z}^2/\lambda_K(H) \leq \|\beta\|_{\Sigma_Z}^2/\xi,$$

where we use (55) in the last step.

■

B.3 Proofs for Section 3.3

B.3.1 Proof of Lemma 6

Let $\bar{A} = A\Sigma_Z^{1/2}$ and $\bar{\beta} \coloneqq \Sigma_Z^{1/2}\beta$. Using $\Sigma_X = A\Sigma_Z A^\top = \bar{A}\bar{A}^\top$, we find

$$\alpha^* = \Sigma_X^+ \bar{A}\bar{\beta} = (\bar{A}\bar{A}^\top)^+ \bar{A}\bar{\beta} = \bar{A}^{+\top}\bar{\beta}, \tag{56}$$

where we use Lemma 32 in the last step. Using this formula, we obtain

$$\|\alpha^*\|_{\Sigma_X}^2 = \bar{\beta}^\top \bar{A}^+ (\bar{A}\bar{A}^\top)\bar{A}^{+\top}\bar{\beta} = \bar{\beta}^\top\bar{\beta} = \|\beta\|_{\Sigma_Z}^2,$$

where we use that $\bar{A}$ is full rank since A and Σ_Z are full rank, and thus $\bar{A}^+\bar{A} = I_K$ by Lemma 32.

Next, by identity (131) in Lemma 32, and the fact that $A^+A = I_K$ and Σ_Z is invertible,

$$\bar{A}^+ = (A\Sigma_Z^{1/2})^+ = \Sigma_Z^{-1/2}A^+.$$

Using this in (56) we find that $\alpha^* = A^{+\top}\beta$, and thus

$$\|\alpha^*\|^2 = \beta^\top A^+ A^{+\top}\beta = \beta^\top (A^\top A)^{-1}A^\top A^{+\top}\beta,$$

where we use $A^+ = (A^\top A)^{-1}A^\top$ by Lemma 32. Thus, again using $A^+A = A^\top A^{+\top} = I_K$, we find

$$\|\alpha^*\|^2 = \beta^\top (A^\top A)^{-1}\beta,$$

as claimed. ■

B.3.2 Proof of Lemma 7

Defining $\bar{A} = A\Sigma_Z^{1/2}$ and $\bar{\beta} = \Sigma_Z^{1/2}\beta$, we have $\alpha^* = \Sigma_X^{-1}\bar{A}\bar{\beta}$. Now recall that since A and Σ_Z are full rank, so is $\bar{A}$ and thus $\bar{A}^+\bar{A} = \bar{A}^\top\bar{A}^{+\top} = I_K$ (see Appendix E). Thus,

$$\begin{aligned}
\alpha^* &= \Sigma_X^{-1}\bar{A}\bar{\beta} \\
&= \Sigma_X^{-1}\bar{A}\bar{A}^\top\bar{A}^{+\top}\bar{\beta} \\
&= \Sigma_X^{-1}(\Sigma_X - \Sigma_E)\bar{A}^{+\top}\bar{\beta} && \text{(since } \Sigma_X = \bar{A}\bar{A}^\top + \Sigma_E\text{)} \\
&= (I_p - \Sigma_X^{-1}\Sigma_E)\bar{A}^{+\top}\bar{\beta}.
\end{aligned}$$

By the Woodbury matrix identity applied to $\Sigma_X^{-1} = (\bar{A}\bar{A}^\top + \Sigma_E)^{-1}$,

$$I_p - \Sigma_X^{-1}\Sigma_E = \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{A}^\top,$$

where $\bar{G} \coloneqq I_K + \bar{A}^\top\Sigma_E^{-1}\bar{A}$. Using this in the previous display,

$$\alpha^* = \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{A}^\top\bar{A}^{+\top}\bar{\beta} = \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{\beta}, \tag{57}$$

where we again use $\bar{A}^+\bar{A} = \bar{A}^\top\bar{A}^{+\top} = I_K$ in the second step.

Bounds on $\|\alpha^*\|^2_{\Sigma_X}$*:* By (57), we find

$$\begin{aligned}\|\alpha^*\|^2_{\Sigma_X} &= \bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top \Sigma_E^{-1}(\bar{A}\bar{A}^\top + \Sigma_E)\Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{\beta} \\ &= \bar{\beta}^\top \bar{G}^{-1}(\bar{A}^\top \Sigma_E^{-1}\bar{A})^2\bar{G}^{-1}\bar{\beta} + \bar{\beta}^\top \bar{G}^{-1}(\bar{A}^\top \Sigma_E^{-1}\bar{A})\bar{G}^{-1}\bar{\beta} \\ &= \bar{\beta}^\top \bar{G}^{-1}(\bar{G} - I_K)^2\bar{G}^{-1}\bar{\beta} + \bar{\beta}^\top \bar{G}^{-1}(\bar{G} - I_K)\bar{G}^{-1}\bar{\beta}.\end{aligned}$$

Expanding the above and simplifying, we find

$$\|\alpha^*\|^2_{\Sigma_X} = \bar{\beta}^\top[I_K - \bar{G}^{-1}]\bar{\beta} = \|\beta\|^2_{\Sigma_Z} - \bar{\beta}^\top \bar{G}^{-1}\bar{\beta}. \tag{58}$$

Recalling that $R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top \bar{G}^{-1}\bar{\beta}$ from (53) above, Lemma 5 implies that

$$0 \le \bar{\beta}^\top \bar{G}^{-1}\bar{\beta} \le \|\beta\|^2_{\Sigma_Z}/\xi.$$

Combining this with (58) yields

$$(1 - \xi^{-1}) \cdot \|\beta\|^2_{\Sigma_Z} \le \|\alpha^*\|^2_{\Sigma_X} \le \|\beta\|^2_{\Sigma_Z}.$$

Thus, when $\xi > c > 1$, $\|\alpha^*\|^2_{\Sigma_X} \asymp \|\beta\|^2_{\Sigma_Z}$, as claimed.

Bounds on $\|\alpha^*\|^2$*:* Using (57), we find

$$\|\alpha^*\|^2 = \bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top \Sigma_E^{-2}\bar{A}\bar{G}^{-1}\bar{\beta}. \tag{59}$$

Thus,

$$\begin{aligned}\|\alpha^*\|^2 &\le \frac{1}{\lambda_p(\Sigma_E)}\bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{\beta} \\ &= \frac{1}{\lambda_p(\Sigma_E)}\bar{\beta}^\top \bar{G}^{-1}(\bar{G} - I_K)\bar{G}^{-1}\bar{\beta} \\ &= \frac{1}{\lambda_p(\Sigma_E)}(\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} - \bar{\beta}^\top \bar{G}^{-2}\bar{\beta}) \\ &\le \frac{1}{\lambda_p(\Sigma_E)}\bar{\beta}^\top \bar{G}^{-1}\bar{\beta}. \qquad (60)\end{aligned}$$

We also have

$$\begin{aligned}\|\alpha^*\|^2 &\ge \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top {\Sigma_E}^{-1}\bar{A}\bar{G}^{-1}\bar{\beta} \\ &= \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}(\bar{G} - I_K)\bar{G}^{-1}\bar{\beta} \\ &= \frac{1}{\|\Sigma_E\|}[\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} - \bar{\beta}^\top \bar{G}^{-2}\bar{\beta}] \\ &\ge \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} \cdot [1 - 1/\lambda_K(\bar{G})] \qquad (61) \\ &\ge \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} \cdot [1 - 1/\xi], \qquad (62)\end{aligned}$$

where in the final step we used

$$\lambda_K(\bar{G}) = 1 + \lambda_K(\bar{A}^\top \Sigma_E^{-1} \bar{A}) \geq \lambda_K(\bar{A}^\top \bar{A})/\|\Sigma_E\| = \xi.$$

Combining (60) and (62),

$$\left(\frac{\xi - 1}{\xi}\right) \frac{1}{\|\Sigma_E\|} \bar{\beta}^\top \bar{G}^{-1} \bar{\beta} \leq \|\alpha^*\|^2 \leq \frac{1}{\lambda_p(\Sigma_E)} \bar{\beta}^\top \bar{G}^{-1} \bar{\beta}.$$

Recalling that $R(\alpha^*) - \sigma_\varepsilon^2 = \bar{\beta}^\top \bar{G}^{-1} \bar{\beta}$ from (53) above, Lemma 5 implies

$$\left(\frac{\xi - 1}{\xi + 1}\right) \frac{1}{\|\Sigma_E\|} \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \leq \|\alpha^*\|^2 \leq \frac{1}{\lambda_p(\Sigma_E)} \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta. \tag{63}$$

As shown at the end of this proof using the singular value decomposition of A, we have that

$$\lambda_p(\Sigma_E) \cdot \beta^\top (A^\top A)^{-1} \beta \leq \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \leq \|\Sigma_E\| \cdot \beta^\top (A^\top A)^{-1} \beta.$$

Combining this with (63) proves that

$$\left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \cdot \beta^\top (A^\top A)^{-1} \beta \leq \|\alpha^*\|^2 \leq \kappa(\Sigma_E) \cdot \beta^\top (A^\top A)^{-1} \beta. \tag{64}$$

Thus, when $\xi > c > 1$ and $\kappa(\Sigma_E) < C$, $\|\alpha^*\|^2 \asymp \beta^\top (A^\top A)^{-1} \beta$, as claimed.

Proof of (64): Write the singular value decomposition $A = U_A S_A V_A^\top$, where U_A is an $p \times K$ matrix with satisfying $U_A^\top U_A = I_K$, V_A is a $K \times K$ orthogonal matrix, and S_A is a $K \times K$ diagonal matrix with positive entries (since we assume $\text{rank}(A) = K$). Then,

$$(A^\top \Sigma_E^{-1} A)^{-1} = (V_A S_A U_A^\top \Sigma_E^{-1} U_A S_A V_A^\top)^{-1} = V_A S_A^{-1} (U_A^\top \Sigma_E^{-1} U_A)^{-1} S_A^{-1} V_A^\top. \tag{65}$$

Thus,

$$\begin{aligned} \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta &= \beta^\top V_A S_A^{-1} (U_A^\top \Sigma_E^{-1} U_A)^{-1} S_A^{-1} V_A^\top \beta \\ &\geq \beta^\top V_A S_A^{-2} V_A^\top \beta \cdot \frac{1}{\|U_A^\top \Sigma_E^{-1} U_A\|}, \end{aligned}$$

so using

$$\|U_A^\top \Sigma_E^{-1} U_A\| \leq \|\Sigma_E^{-1}\| = 1/\lambda_p(\Sigma_E),$$

we find

$$\beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \geq \lambda_p(\Sigma_E) \cdot \beta^\top V_A S_A^{-2} V_A^\top \beta. \tag{66}$$

We next observe that since $U_A^\top U_A = I_K$

$$(A^\top A)^{-1} = (V_A S_A U_A^\top U_A S_A V_A^\top)^{-1} = V_A S_A^{-2} V_A^\top, \tag{67}$$

and thus, by (66),

$$\beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \geq \lambda_p(\Sigma_E) \cdot \beta^\top (A^\top A)^{-1} \beta,$$

which proves the lower bound in (64). To prove the upper bound, we use that by (65),

$$\begin{aligned}\beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta &= \beta^\top V_A S_A^{-1} (U_A^\top \Sigma_E^{-1} U_A)^{-1} S_A^{-1} V_A^\top \beta \\ &\le \beta^\top V_A S_A^{-2} V_A^\top \beta \cdot \frac{1}{\lambda_K(U_A^\top \Sigma_E^{-1} U_A)}.\end{aligned}$$

Thus, since

$$\lambda_K(U_A^\top \Sigma_E^{-1} U_A) \ge \lambda_K(U_A^\top U_A)\lambda_p(\Sigma_E^{-1}) = 1/\|\Sigma_E\|,$$

we have

$$\beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \le \|\Sigma_E\| \cdot \beta^\top V_A S_A^{-2} V_A^\top \beta = \|\Sigma_E\| \cdot \beta^\top (A^\top A)^{-1} \beta,$$

where in the last step we use (67). This establishes the upper bound of (64), completing the proof.

■

B.3.3 Proof of Corollary 9

Under the conditions stated, by either Lemma 6 or Lemma 7, $\|\alpha^*\|^2 \lesssim \beta^\top (A^\top A)^{-1} \beta$. Thus, using that Σ_Z is invertible,

$$\|\alpha^*\|^2 \lesssim \beta^\top (A^\top A)^{-1} \beta = \beta^\top \Sigma_Z^{1/2} (\Sigma_Z^{1/2} A^\top A \Sigma_Z^{1/2})^{-1} \Sigma_Z^{1/2} \beta \le \|\beta\|_{\Sigma_Z}^2 / \lambda_K(A \Sigma_Z A^\top), \quad (68)$$

so $\|\alpha^*\| \to 0$ when $\|\beta\|_{\Sigma_Z}^2 / \lambda_K(A\Sigma_Z A^\top) \to 0$.

For the second claim, we have

$$\begin{aligned}R(\mathbf{0}) - R(\alpha^*) &= \|\alpha^*\|_{\Sigma_X}^2 && \text{(by (16))} \\ &\gtrsim \|\beta\|_{\Sigma_Z}^2. && \text{(by either Lemma 6 or Lemma 7)}\end{aligned}$$

The claim follows by taking the limit inferior as $p \to \infty$ on both sides of the inequality and using condition (15). ■

Appendix C. Proofs for Section 4

C.1 Proofs for Section 4.1

In the proofs of Lemma 11 and Theorem 12, we will use the event

$$\mathcal{A} := \left\{ \|\tilde{\mathbf{Z}}^+ \tilde{\boldsymbol{\varepsilon}}\|^2 \lesssim \log(n) \mathrm{tr}(\tilde{\mathbf{Z}}^{+\top} \tilde{\mathbf{Z}}^+),\ c_1 n \le \sigma_K^2(\tilde{\mathbf{Z}}) \le \|\tilde{\mathbf{Z}}\|^2 \le c_2 n \right\}, \quad (69)$$

which occurs with probability at least $1 - c/n$, as shown in Lemma 22 below, where $\mathbf{Z} = \tilde{\mathbf{Z}} \Sigma_Z^{1/2}$ and $\boldsymbol{\varepsilon} = \sigma_\varepsilon \tilde{\boldsymbol{\varepsilon}}$ by Assumption 3.

C.1.1 Proof of Lemma 11

On the event $\mathcal{A}$ defined in (69), and using $\lambda_K(\Sigma_Z) > 0$ by Assumption 2,

$$\sigma_K^2(\mathbf{Z}) = \lambda_K(\mathbf{Z}\mathbf{Z}^\top) = \lambda_K(\tilde{\mathbf{Z}} \Sigma_Z \tilde{\mathbf{Z}}^\top) \ge \lambda_K(\Sigma_Z) \cdot \sigma_n^2(\tilde{\mathbf{Z}}) \gtrsim \lambda_K(\Sigma_Z) \cdot n > 0, \quad (70)$$

so $\text{rank}(\mathbf{Z}) = K$ and thus $\mathbf{Z}^+\mathbf{Z} = I_K$ by Lemma 32 in Appendix E. Similarly, since A is of dimension $p \times K$ and $\text{rank}(A) = K$ by Assumption 2,

$$A^\top A^{+\top} = (A^+ A)^\top = I_K.$$

Using these two results together with (131) of Lemma 32, we find

$$\mathbf{X}^+ = (\mathbf{Z}A^\top)^+ = (\mathbf{Z}^+\mathbf{Z}A^\top)^+(\mathbf{Z}A^\top A^{+\top})^+ = A^{+\top}\mathbf{Z}^+. \tag{71}$$

Thus, on the event $\mathcal{A}$,

$$\widehat{\alpha} = \mathbf{X}^+\mathbf{y} = A^{+\top}\mathbf{Z}^+\mathbf{y}, \tag{72}$$

so

$$\begin{aligned}
\|\widehat{\alpha}\|^2 &= \|A^{+\top}\mathbf{Z}^+\mathbf{y}\|^2 \\
&= \|A^{+\top}\mathbf{Z}^+\mathbf{Z}\beta + A^{+\top}\mathbf{Z}^+\boldsymbol{\varepsilon}\|^2 && (\text{by } \mathbf{y} = \mathbf{Z}\beta + \boldsymbol{\varepsilon}) \\
&\le 2\|A^{+\top}\beta\|^2 + 2\|A^{+\top}\mathbf{Z}^+\boldsymbol{\varepsilon}\|^2 && (\text{since } \mathbf{Z}^+\mathbf{Z} = I_K \text{ on } \mathcal{A}) \\
&= 2\|A^{+\top}\beta\|^2 + 2\|(A\Sigma_Z^{1/2})^{+\top}\tilde{\mathbf{Z}}^+\boldsymbol{\varepsilon}\|^2,
\end{aligned}$$

where in the last step we used that by Lemma 32,

$$A^{+\top}\mathbf{Z} = A^{+\top}(\tilde{\mathbf{Z}}\Sigma_Z^{1/2})^+ = A^{+\top}\Sigma_Z^{-1/2}\tilde{\mathbf{Z}}^+ = (A\Sigma_Z^{1/2})^{+\top}\tilde{\mathbf{Z}}^+.$$

Continuing, and using

$$A^+A^{+\top} = (A^\top A)^{-1}A^\top A^{+\top} = (A^\top A)^{-1},$$

we find

$$\begin{aligned}
\|\widehat{\alpha}\|^2 &\lesssim \beta^\top(A^\top A)^{-1}\beta + \|(A\Sigma_Z^{1/2})^+\|^2 \cdot \sigma_\varepsilon^2 \cdot \|\tilde{\mathbf{Z}}^+\tilde{\boldsymbol{\varepsilon}}\|^2 \\
&\lesssim \beta^\top(A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma_\varepsilon^2\log(n)\text{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^+) && (\text{on } \mathcal{A}) \\
&\le \beta^\top(A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma_\varepsilon^2\log(n)K\|\tilde{\mathbf{Z}}^+\|^2 \\
&= \beta^\top(A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma_\varepsilon^2\log(n)K\frac{1}{\sigma_K^2(\tilde{\mathbf{Z}})} \\
&\lesssim \beta^\top(A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma_\varepsilon^2\log(n)\frac{K}{n} && (\text{on } \mathcal{A}) \\
&\le \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\left(\|\beta\|_{\Sigma_Z}^2 + \sigma_\varepsilon^2\log(n)\frac{K}{n}\right). && (\text{by (68)})
\end{aligned}$$

Under the assumptions of this Lemma, the event $\mathcal{A}$ holds with probability at least $1 - c/n$ by Lemma 22, so the proof is complete. ■

C.1.2 Proof of Theorem 12

Part 1: By (72), $\widehat{\alpha} = A^{+\top}\mathbf{Z}^{+}\mathbf{y}$ on the event $\mathcal{A}$ defined in (69). Thus, using $X = AZ$ and $A^{\top}A^{+\top} = I_K$ since A is full rank by Assumption 2,

$$\widehat{y}_x = X^{\top}\widehat{\alpha} = Z^{\top}A^{\top}A^{+\top}\mathbf{Z}^{+}\mathbf{y} = Z^{\top}\mathbf{Z}^{+}\mathbf{y} = Z^{\top}\widehat{\beta} = \widehat{y}_z. \tag{73}$$

Part 2: Using the independence of ε and Z together with (73), the excess risk can be written as

$$R(\widehat{\alpha}) - \sigma_\varepsilon^2 = \mathbb{E}[(X^{\top}\widehat{\alpha} - Z^{\top}\beta)^2] = \mathbb{E}[(Z^{\top}\widehat{\beta} - Z^{\top}\beta)^2] = \|\widehat{\beta} - \beta\|_{\Sigma_Z}^2. \tag{74}$$

By (70), $\operatorname{rank}(\mathbf{Z}) = K$ and $\mathbf{Z}^{+}\mathbf{Z} = I_K$ on the event $\mathcal{A}$ defined in (69). Thus,

$$\widehat{\beta} = \mathbf{Z}^{+}\mathbf{y} = \mathbf{Z}^{+}\mathbf{Z}\beta + \mathbf{Z}^{+}\boldsymbol{\varepsilon} = \beta + \mathbf{Z}^{+}\boldsymbol{\varepsilon},$$

so by (74),

$$R(\widehat{\alpha}) - \sigma_\varepsilon^2 = \|\mathbf{Z}^{+}\boldsymbol{\varepsilon}\|_{\Sigma_Z}^2 = \|\Sigma_Z^{1/2}\mathbf{Z}^{+}\boldsymbol{\varepsilon}\|^2. \tag{75}$$

By (131) of Lemma 32,

$$\Sigma_Z^{1/2}\mathbf{Z}^{+} = \Sigma_Z^{1/2}(\tilde{\mathbf{Z}}\Sigma_Z^{1/2})^{+} = \Sigma_Z^{1/2}(\tilde{\mathbf{Z}}^{+}\tilde{\mathbf{Z}}\Sigma_Z^{1/2})^{+}(\tilde{\mathbf{Z}}\Sigma_Z^{1/2}\Sigma_Z^{-1/2})^{+} = \Sigma_Z^{1/2}\Sigma_Z^{-1/2}\tilde{\mathbf{Z}}^{+} = \tilde{\mathbf{Z}}^{+}, \tag{76}$$

where we used that $\tilde{\mathbf{Z}}^{+}\tilde{\mathbf{Z}} = I_K$ since $\operatorname{rank}(\tilde{\mathbf{Z}}) = K$ on $\mathcal{A}$. Thus by (75), we find that on $\mathcal{A}$,

$$R(\widehat{\alpha}) - \sigma_\varepsilon^2 = \|\tilde{\mathbf{Z}}^{+}\boldsymbol{\varepsilon}\|^2 = \sigma_\varepsilon^2\|\tilde{\mathbf{Z}}^{+}\tilde{\boldsymbol{\varepsilon}}\|^2 \lesssim \sigma_\varepsilon^2 \log(n)\operatorname{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^{+}). \tag{77}$$

We then use that $\operatorname{rank}(\tilde{\mathbf{Z}}^{+}) = K$ and that $\|\tilde{\mathbf{Z}}^{+}\| = 1/\sigma_K(\tilde{\mathbf{Z}})$ from Lemma 32 in Appendix E below to find that on $\mathcal{A}$,

$$\operatorname{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^{+}) \le K\|\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^{+}\| = K\|\tilde{\mathbf{Z}}^{+}\|^2 = \frac{K}{\sigma_K^2(\tilde{\mathbf{Z}})} \lesssim \frac{K}{n}.$$

Plugging this into (77) completes the proof of the upper bound.

For the lower bound, first observe that on $\mathcal{A}$,

$$\mathbb{E}_{\boldsymbol{\varepsilon}}R(\widehat{\alpha}) - \sigma_\varepsilon^2 = \mathbb{E}_{\boldsymbol{\varepsilon}}\|\tilde{\mathbf{Z}}^{+}\boldsymbol{\varepsilon}\|^2 = \sigma_\varepsilon^2\operatorname{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^{+}) \ge \sigma_\varepsilon^2 K\lambda_K(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^{+}) = \sigma_\varepsilon^2 K\sigma_K^2(\tilde{\mathbf{Z}}^{+}),$$

so using $\sigma_K(\tilde{\mathbf{Z}}^{+}) = 1/\|\tilde{\mathbf{Z}}\|$ by Lemma 32 again,

$$\mathbb{E}_{\boldsymbol{\varepsilon}}R(\widehat{\alpha}) - \sigma_\varepsilon^2 \ge \sigma_\varepsilon^2\frac{K}{\|\tilde{\mathbf{Z}}\|^2} \gtrsim \sigma_\varepsilon^2\frac{K}{n}.$$

■

Lemma 22. *Suppose that Assumptions 2 & 3 hold and that $n > C \cdot K$ for some large enough absolute constant $C > 0$. Then there exists $c > 0$ such that*

$$\mathbb{P}\left\{\|\tilde{\mathbf{Z}}^{+}\tilde{\boldsymbol{\varepsilon}}\|^2 \lesssim \log(n)\operatorname{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^{+}),\ c_1 n \le \sigma_K^2(\tilde{\mathbf{Z}}) \le \|\tilde{\mathbf{Z}}\|^2 \le c_2 n\right\} \ge 1 - c/n.$$

Proof Since $\tilde{\mathbf{Z}}$ has independent rows with entries that are zero mean, unit variance, and have sub-Gaussian constants bounded by an absolute constant, Theorem 4.6.1 of Vershynin (2019) gives that with probability at least $1 - 2/n$,

$$\sqrt{n} - c''(\sqrt{K} + \sqrt{\log n}) \leq \sigma_n(\tilde{\mathbf{Z}}) \leq \|\tilde{\mathbf{Z}}\| \leq \sqrt{n} + c''(\sqrt{K} + \sqrt{\log n}).$$

and thus

$$\sqrt{n} \cdot [1 - c''(\sqrt{K/n} + \sqrt{\log(n)/n})] \leq \sigma_n(\tilde{\mathbf{Z}}) \leq \|\tilde{\mathbf{Z}}\| \leq \sqrt{n} \cdot [1 - c''(\sqrt{K/n} + \sqrt{\log(n)/n})].$$

Using that $n > CK$ we can choose C large enough such that

$$c''(\sqrt{K/n} + \sqrt{\log(n)/n}) < c_0 < 1,$$

and thus

$$\mathbb{P}\left(c_3 n \leq \sigma_K^2(\tilde{\mathbf{Z}}) \leq \|\tilde{\mathbf{Z}}\|^2 \leq c_4 n\right) \geq 1 - 2/n. \tag{78}$$

The bound

$$\mathbb{P}\left(\|\tilde{\mathbf{Z}}^+ \tilde{\boldsymbol{\varepsilon}}\|^2 \lesssim \log(n) \mathrm{tr}[\tilde{\mathbf{Z}}^{+\top} \tilde{\mathbf{Z}}^+]\right) \geq 1 - e^{-cn}$$

follows from Lemma 23, which we state below. Combining this with (78) proves that $\mathcal{A}$ occurs with probability at least $1 - c/n$. ■

The following result is a slightly adapted version of Lemma 19 from Bartlett et al. (2020) and the discussion that follows.

Lemma 23. *Suppose $\tilde{\boldsymbol{\varepsilon}} \in \mathbb{R}^n$ has independent entries with sub-Gaussian constants bounded by an absolute constant, and suppose $M \in \mathbb{R}^{n \times n}$ is a positive semidefinite matrix independent of $\tilde{\boldsymbol{\varepsilon}}$. Then, with probability at least $1 - e^{-cn}$,*

$$\tilde{\boldsymbol{\varepsilon}}^\top M \tilde{\boldsymbol{\varepsilon}} \lesssim \log(n) \cdot \mathrm{tr}(M).$$

C.1.3 Proof of Lemma 13

Suppose $\mathrm{rank}(\mathbf{X}) = K$. We can then write the singular value decomposition of $\mathbf{X}$ as $\mathbf{X} = \widehat{V}_K \widehat{D} \widehat{U}_K^\top$, where $\widehat{V}_K \in \mathbb{R}^{n \times K}$, $\widehat{U}_K \in \mathbb{R}^{p \times K}$, and $\widehat{D} \in \mathbb{R}^{K \times K}$ are full rank, and $\widehat{V}_K^\top \widehat{V}_K = \widehat{U}_K^\top \widehat{U}_K = I_K$. Thus,

$$(\mathbf{X}\widehat{U}_K)^+ = (\widehat{V}_K \widehat{D} \widehat{U}_K^\top \widehat{U}_K)^+ = (\widehat{V}_K \widehat{D})^+.$$

By Lemma 32 of Appendix E, we thus have

$$\begin{aligned}
(\mathbf{X}\widehat{U}_K)^+ &= (\widehat{V}_K^+ \widehat{V}_K \widehat{D})^+ (\widehat{V}_K \widehat{D} \widehat{D}^+)^+ \\
&= \widehat{D}^+ \widehat{V}_K^+ && \text{(since } \widehat{V}_K \text{ and } \widehat{D} \text{ full rank)} \\
&= \widehat{D}^+ (\widehat{V}_K^\top \widehat{V}_K)^+ \widehat{V}_K^\top \\
&= \widehat{D}^+ \widehat{V}_K^\top. && \text{(by } \widehat{V}_K^\top \widehat{V}_K = I_K)
\end{aligned}$$

We thus find

$$\widehat{\alpha}_{\mathrm{PCR}} = \widehat{U}_K (\mathbf{X}\widehat{U}_K)^+ \mathbf{y} = \widehat{U}_K \widehat{D}^+ \widehat{V}_K^\top \mathbf{y} = \mathbf{X}^+ \mathbf{y} = \widehat{\alpha},$$

where we recognize $\widehat{U}_K \widehat{D}^+ \widehat{V}_K^\top$ as the pseudoinverse of $\mathbf{X}$ in the third step.

Now suppose that Assumptions 2 & 3 hold and $K > C \cdot n$. Then by Lemma 22 above, $\mathbb{P}\{\sigma_K^2(\tilde{\mathbf{Z}}) \gtrsim n\} \geq 1 - c/n$. Thus, using

$$\sigma_K^2(\mathbf{Z}) = \sigma_K^2(\tilde{\mathbf{Z}}\Sigma_Z^{1/2}) \geq \lambda_K(\Sigma_Z)\sigma_K^2(\tilde{\mathbf{Z}})$$

and that $\lambda_K(\Sigma_Z) > 0$ by Assumption 2,

$$\mathbb{P}\{\text{rank}(\mathbf{X}) = K\} \geq \mathbb{P}\{\sigma_K^2(\mathbf{Z}) \gtrsim n\} \geq \mathbb{P}\{\sigma_K^2(\tilde{\mathbf{Z}}) \gtrsim n\} \geq 1 - c/n,$$

which completes the proof. ■

C.2 Proofs for Section 4.2

In this section we begin with the proof of Lemma 15 and our main result, Theorem 16, which rely on Proposition 14, proved subsequently. The proofs of Lemma 15 and Theorem 16 use the event

$$\mathcal{E} \coloneqq \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3, \tag{79}$$

where for positive absolute constants c_1 to c_6,

$$\mathcal{E}_1 \coloneqq \left\{\sigma_n^2(\mathbf{X}) \geq c_1 \text{tr}(\Sigma_E), \|\mathbf{E}\|^2 \leq c_2 \text{tr}(\Sigma_E),\ c_3 n \leq \sigma_K^2(\tilde{\mathbf{Z}}) \leq \|\tilde{\mathbf{Z}}\|^2 \leq c_4 n\right\},$$

$$\mathcal{E}_2 \coloneqq \left\{\tilde{\boldsymbol{\varepsilon}}^\top \mathbf{X}^{+\top} \Sigma_X \mathbf{X}^+ \tilde{\boldsymbol{\varepsilon}} \leq c_5 \log(n) \text{tr}(\mathbf{X}^{+\top} \Sigma_X \mathbf{X}^+)\right\},$$

$$\mathcal{E}_3 \coloneqq \left\{\tilde{\boldsymbol{\varepsilon}}^\top \mathbf{X}^{+\top} \mathbf{X}^+ \tilde{\boldsymbol{\varepsilon}} \leq c_6 \log(n) \text{tr}(\mathbf{X}^{+\top} \mathbf{X}^+)\right\}.$$

We will show in Lemma 24 below that $\mathcal{E}$ occurs with probability at least $1 - c/n$ for an absolute constant $c > 0$.

C.2.1 Proof of Theorem 15

Using $\widehat{\alpha} = \mathbf{X}^+\mathbf{y}$, $\mathbf{y} = \mathbf{Z}\beta + \boldsymbol{\varepsilon}$, and that A is full rank by Assumption 2, we find

$$\begin{aligned}
\widehat{\alpha} &= \mathbf{X}^+\mathbf{y} \\
&= \mathbf{X}^+\mathbf{Z}\beta + \mathbf{X}^+\boldsymbol{\varepsilon} \\
&= \mathbf{X}^+\mathbf{Z}A^\top A^{+\top}\beta + \mathbf{X}^+\boldsymbol{\varepsilon} && (A^+A = I_K \text{ since } \text{rank}(A) = K) \\
&= \mathbf{X}^+(\mathbf{X} - \mathbf{E})A^{+\top}\beta + \mathbf{X}^+\boldsymbol{\varepsilon} && (\text{using } \mathbf{X} = \mathbf{Z}A^\top + \mathbf{E}) \\
&= \mathbf{X}^+\mathbf{X}A^{+\top}\beta - \mathbf{X}^+\mathbf{E}A^{+\top}\beta + \mathbf{X}^+\boldsymbol{\varepsilon}.
\end{aligned}$$

Thus, using $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$,

$$\begin{aligned}
\|\widehat{\alpha}\|^2 &\leq 3\|\mathbf{X}^+\mathbf{X}A^{+\top}\beta\|^2 + 3\|\mathbf{X}^+\mathbf{E}A^{+\top}\beta\|^2 + 3\|\mathbf{X}^+\boldsymbol{\varepsilon}\|^2 \\
&\lesssim \|\mathbf{X}^+\mathbf{X}\|^2\|A^{+\top}\beta\|^2 + \frac{\|\mathbf{E}\|^2}{\sigma_n^2(\mathbf{X})}\|A^{+\top}\beta\|^2 + \sigma_\varepsilon^2 \tilde{\boldsymbol{\varepsilon}}^\top \mathbf{X}^{+\top}\mathbf{X}^+\tilde{\boldsymbol{\varepsilon}} \\
&\leq \|A^{+\top}\beta\|^2 + \|A^{+\top}\beta\|^2 + \sigma_\varepsilon^2 \log(n)\text{tr}(\mathbf{X}^{+\top}\mathbf{X}^+),
\end{aligned}$$

where in the last step holds on the event $\mathcal{E}$, and uses that $\|\mathbf{X}^+\mathbf{X}\| \leq 1$ since $\mathbf{X}^+\mathbf{X}$ is a projection matrix. Recalling that by (68),

$$\|A^{+\top}\beta\|^2 = \beta^\top (A^\top A)^{-1}\beta \leq \|\beta\|^2_{\Sigma_Z}/\lambda_K(A\Sigma_Z A^\top),$$

and using that $\operatorname{rank}(\mathbf{X}) \leq n$, we find that on $\mathcal{E}$,

$$\begin{aligned}
\|\widehat{\alpha}\|^2 &\lesssim \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\|\beta\|^2_{\Sigma_Z} + \sigma_\varepsilon^2 \log(n)\cdot n \cdot \|\mathbf{X}^+\|^2 \\
&= \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\|\beta\|^2_{\Sigma_Z} + \sigma_\varepsilon^2 \frac{n\log n}{\sigma_n^2(\mathbf{X})} \\
&\lesssim \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\|\beta\|^2_{\Sigma_Z} + \sigma_\varepsilon^2 \frac{n\log n}{\operatorname{tr}(\Sigma_E)}.
\end{aligned}$$

By Lemma 24, $\mathcal{E}$ holds with probability at least $1 - c/n$, so the proof is complete. ■

C.2.2 Proof of Theorem 16

Using that Z, E and ε are independent of one another and of $\widehat{\alpha}$, we have

$$\begin{aligned}
R(\widehat{\alpha}) &= \mathbb{E}[(X^\top\widehat{\alpha} - y)^2] \\
&= \mathbb{E}[(Z^\top A^\top\widehat{\alpha} - Z^\top\beta - \varepsilon + E^\top\widehat{\alpha})^2] \\
&= \sigma_\varepsilon^2 + \|\Sigma_E^{1/2}\widehat{\alpha}\|^2 + \|\Sigma_Z^{1/2}(A^\top\widehat{\alpha} - \beta)\|^2.
\end{aligned}$$

Since $\widehat{\alpha} = \mathbf{X}^+\mathbf{y} = \mathbf{X}^+\mathbf{Z}\beta + \mathbf{X}^+\boldsymbol{\varepsilon}$,

$$\|\Sigma_E^{1/2}\widehat{\alpha}\|^2 \leq 2\|\Sigma_E^{1/2}\mathbf{X}^+\mathbf{Z}\beta\|^2 + 2\|\Sigma_E^{1/2}\mathbf{X}^+\boldsymbol{\varepsilon}\|^2 \coloneqq 2B_1 + 2V_1.$$

Similarly,

$$\|\Sigma_Z^{1/2}(A^\top\widehat{\alpha} - \beta)\|^2 \leq 2\|\Sigma_Z^{1/2}(A^\top\mathbf{X}^+\mathbf{Z} - I_K)\beta\|^2 + 2\|\Sigma_Z^{1/2}A^\top\mathbf{X}^+\boldsymbol{\varepsilon}\|^2 \coloneqq 2B_2 + 2V_2.$$

We thus have $R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim B + V$, where we view $B \coloneqq B_1 + B_2$ as a bound on the bias component of the risk and $V \coloneqq V_1 + V_2$ as a bound on the variance component. In what follows, we bound the four terms

$$\begin{aligned}
B_1 &= \|\Sigma_E^{1/2}\mathbf{X}^+\mathbf{Z}\beta\|^2 \\
B_2 &= \|\Sigma_Z^{1/2}(A^\top\mathbf{X}^+\mathbf{Z} - I_K)\beta\|^2 \\
V_1 &= \|\Sigma_E^{1/2}\mathbf{X}^+\boldsymbol{\varepsilon}\|^2 \\
V_2 &= \|\Sigma_Z^{1/2}A^\top\mathbf{X}^+\boldsymbol{\varepsilon}\|^2.
\end{aligned}$$

Bounding the bias component: On the event $\mathcal{E}$ defined in (79), $\sigma_n(\mathbf{X}) > 0$ and by Assumption 2 and (70) above, $\sigma_n^2(\mathbf{Z}) \gtrsim \lambda_K(\Sigma_Z)n > 0$. Thus $\mathbf{X}$ and $\mathbf{Z}$ are of rank n and K respectively, so by Lemma 32 of Appendix E, $\mathbf{X}\mathbf{X}^+ = I_n$ and $\mathbf{Z}^+\mathbf{Z} = I_K$. It follows that

$$\begin{aligned}
\mathbf{Z}^+ - A^\top\mathbf{X}^+ &= \mathbf{Z}^+\mathbf{X}\mathbf{X}^+ - A^\top\mathbf{X}^+ && (\text{since } \mathbf{X}\mathbf{X}^+ = I_n) \\
&= (\mathbf{Z}^+\mathbf{X} - A^\top)\mathbf{X}^+ \\
&= (\mathbf{Z}^+[\mathbf{Z}A^\top + \mathbf{E}] - A^\top)\mathbf{X}^+ && (\text{since } \mathbf{X} = \mathbf{Z}A^\top + \mathbf{E}) \\
&= \mathbf{Z}^+\mathbf{E}\mathbf{X}^+, && (\text{since } \mathbf{Z}^+\mathbf{Z} = I_K) \qquad (80)
\end{aligned}$$

and thus again using $\mathbf{Z}^+\mathbf{Z} = I_K$

$$B_2 = \|\Sigma_Z^{1/2}(A^\top \mathbf{X}^+\mathbf{Z} - I_K)\beta\|^2 = \|\Sigma_Z^{1/2}(A^\top \mathbf{X}^+ - \mathbf{Z}^+)\mathbf{Z}\beta\|^2 = \|\Sigma_Z^{1/2}\mathbf{Z}^+\mathbf{E}\mathbf{X}^+\mathbf{Z}\beta\|^2.$$

By (76) above and the fact that $\mathbf{Z}$ is full rank on $\mathcal{E}$, $\Sigma_Z^{1/2}\mathbf{Z}^+ = \tilde{\mathbf{Z}}^+$, so on $\mathcal{E}$,

$$B_2 = \|\tilde{\mathbf{Z}}^+\mathbf{E}\mathbf{X}^+\mathbf{Z}\beta\|^2 \leq \frac{\|\mathbf{E}\|^2}{\sigma_K^2(\tilde{\mathbf{Z}})}\|\mathbf{X}^+\mathbf{Z}\beta\|^2 \lesssim \frac{\mathrm{tr}(\Sigma_E)\|\mathbf{X}^+\mathbf{Z}\beta\|^2}{n},$$

where we also used that $\|\tilde{\mathbf{Z}}^+\|^2 = 1/\sigma_K^2(\tilde{\mathbf{Z}})$. Since $B_1 = \|\Sigma_E^{1/2}\mathbf{X}^+\mathbf{Z}\beta\|^2 \leq \|\Sigma_E\|\|\mathbf{X}^+\mathbf{Z}\beta\|^2$, and

$$\|\Sigma_E\| = \mathrm{tr}(\Sigma_E)\frac{\|\Sigma_E\|}{\mathrm{tr}(\Sigma_E)} = \frac{\mathrm{tr}(\Sigma_E)}{n}\cdot\frac{n}{\mathrm{r_e}(\Sigma_E)} \lesssim \frac{\mathrm{tr}(\Sigma_E)}{n},$$

where we used the assumption $\mathrm{r_e}(\Sigma_E) > c_1 n$ in the last step, we also have that on $\mathcal{E}$,

$$B = B_1 + B_2 \lesssim \frac{\mathrm{tr}(\Sigma_E)\|\mathbf{X}^+\mathbf{Z}\beta\|^2}{n}. \tag{81}$$

To bound $\|\mathbf{X}^+\mathbf{Z}\beta\|^2$, we first use $A^\top A^{+\top} = I_K$ and $\mathbf{Z}A^\top = \mathbf{X} - \mathbf{E}$ to find

$$\|\mathbf{X}^+\mathbf{Z}\beta\|^2 = \|\mathbf{X}^+\mathbf{Z}A^\top A^{+\top}\beta\|^2 \leq 2\|\mathbf{X}^+\mathbf{X}A^{+\top}\beta\|^2 + 2\|\mathbf{X}^+\mathbf{E}A^{+\top}\beta\|^2.$$

The second term can be bounded, on the event $\mathcal{E}$, by

$$\frac{\|\mathbf{E}\|^2\|A^{+\top}\beta\|^2}{\sigma_n^2(\mathbf{X})} \lesssim \|A^{+\top}\beta\|^2.$$

On the other hand, the first term can be bounded as $\|\mathbf{X}^+\mathbf{X}A^{+\top}\beta\|^2 \leq \|A^{+\top}\beta\|^2$ using the fact that $\mathbf{X}^+\mathbf{X}$ is a projection matrix, so we find that on $\mathcal{E}$,

$$\|\mathbf{X}^+\mathbf{Z}\beta\|^2 \lesssim \|A^{+\top}\beta\|^2. \tag{82}$$

Finally, we have

$$\|A^{+\top}\beta\|^2 = \beta^\top(A^\top A)^{-1}\beta = \beta^\top\Sigma_Z^{1/2}(\Sigma_Z^{1/2}A^\top A\Sigma_Z^{1/2})^{-1}\Sigma_Z^{1/2}\beta \leq \frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)}. \tag{83}$$

Combining this with (82) and plugging into (81), we find that on the event $\mathcal{E}$,

$$B \lesssim \frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)}\frac{\mathrm{tr}(\Sigma_E)}{n} = \frac{\|\beta\|_{\Sigma_Z}^2\|\Sigma_E\|}{\lambda_K(A\Sigma_Z A^\top)}\cdot\frac{\mathrm{tr}(\Sigma_E)}{\|\Sigma_E\|n} = \frac{\|\beta\|_{\Sigma_Z}^2}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}. \tag{84}$$

Bounding the variance component: First note that

$$V = V_1 + V_2 = \|\Sigma_E^{1/2}\mathbf{X}^+\boldsymbol{\varepsilon}\|^2 + \|\Sigma_Z^{1/2}A^\top\mathbf{X}^+\boldsymbol{\varepsilon}\|^2 = \boldsymbol{\varepsilon}^\top\mathbf{X}^{+\top}\Sigma_X\mathbf{X}^+\boldsymbol{\varepsilon} = \sigma_\varepsilon^2\tilde{\boldsymbol{\varepsilon}}\mathbf{X}^{+\top}\Sigma_X\mathbf{X}^+\tilde{\boldsymbol{\varepsilon}},$$

so on the event $\mathcal{E}$,

$$V \lesssim \sigma_\varepsilon^2\log(n)\mathrm{tr}(\mathbf{X}^{+\top}\Sigma_X\mathbf{X}^+) = \sigma_\varepsilon^2\log(n)\left\{\mathrm{tr}(\mathbf{X}^{+\top}\Sigma_E\mathbf{X}^+) + \mathrm{tr}(\mathbf{X}^{+\top}A\Sigma_Z A^\top\mathbf{X}^+)\right\}, \tag{85}$$

where we use $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$ in the second step. The first term in (85) can by bounded as

$$\mathrm{tr}(\mathbf{X}^{+\top}\Sigma_E\mathbf{X}^+) \le \|\Sigma_E\| \cdot n\|\mathbf{X}^{+\top}\mathbf{X}^+\| = \|\Sigma_E\|\frac{n}{\sigma_n^2(\mathbf{X})} \lesssim \frac{n}{\mathrm{r_e}(\Sigma_E)}, \tag{86}$$

where in the first step we used that $\mathrm{rank}(\mathbf{X}^+) = \mathrm{rank}(\mathbf{X}) = n$ and in the last step that $\sigma_n^2(\mathbf{X}) \gtrsim \mathrm{tr}(\Sigma_E)$ on $\mathcal{E}$.

For the second term in (85),

$$\begin{aligned}
\mathrm{tr}(\mathbf{X}^{+\top}A\Sigma_Z A^\top\mathbf{X}^+) &\le K\|\Sigma_Z^{1/2}A^\top\mathbf{X}^+\|^2 && (\text{since } \mathrm{rank}(A\Sigma_Z A^\top) = K)\\
&= K\|\Sigma_Z^{1/2}(\mathbf{Z}^+ - \mathbf{Z}^+\mathbf{E}\mathbf{X}^+)\|^2 && (\text{by (80) above})\\
&\le 2K\|\tilde{\mathbf{Z}}^+\|^2 + 2K\|\tilde{\mathbf{Z}}^+\|^2\|\mathbf{E}\|^2\|\mathbf{X}^+\|^2,
\end{aligned}$$

where we use that $\Sigma_Z^{1/2}\mathbf{Z}^+ = \tilde{\mathbf{Z}}^+$ from (76) in the final step. Continuing, we find

$$\mathrm{tr}(\mathbf{X}^{+\top}A\Sigma_Z A^\top\mathbf{X}^+) \lesssim \frac{K}{\sigma_K^2(\tilde{\mathbf{Z}})}\left(1 + \frac{\|\mathbf{E}\|^2}{\sigma_n^2(\mathbf{X})}\right) \lesssim \frac{K}{n}, \tag{87}$$

where we use the bounds defining $\mathcal{E}_1$ in the last inequality. Combining (87) and (86) with (85), we conclude that on $\mathcal{E}$,

$$V \lesssim \sigma_\varepsilon^2\frac{n\log n}{\mathrm{r_e}(\Sigma_E)} + \sigma_\varepsilon^2\frac{K\log n}{n}.$$

Combining this with the bias bound (84) gives the bound in the statement of the theorem. By Lemma 24 below, $\mathbb{P}(\mathcal{E}) \ge 1 - c/n$, so the proof is complete. ■

Lemma 24. *Under model (5), suppose that Assumptions 2 and 3 hold and $n > C \cdot K$ and $\mathrm{r_e}(\Sigma_E) > C \cdot n$ hold, for some $C > 0$. Then $\mathbb{P}(\mathcal{E}) \ge 1 - c/n$, where $\mathcal{E} \coloneqq \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and*

$$\mathcal{E}_1 \coloneqq \left\{\sigma_n^2(\mathbf{X}) \ge c_1\mathrm{tr}(\Sigma_E), \|\mathbf{E}\|^2 \le c_2\mathrm{tr}(\Sigma_E),\ c_3 n \le \sigma_K^2(\tilde{\mathbf{Z}}) \le \|\tilde{\mathbf{Z}}\|^2 \le c_4 n\right\},$$

$$\mathcal{E}_2 \coloneqq \left\{\tilde{\boldsymbol{\varepsilon}}^\top\mathbf{X}^{+\top}\Sigma_X\mathbf{X}^+\tilde{\boldsymbol{\varepsilon}} \le c_5\log(n)\mathrm{tr}(\mathbf{X}^{+\top}\Sigma_X\mathbf{X}^+)\right\},$$

$$\mathcal{E}_3 \coloneqq \left\{\tilde{\boldsymbol{\varepsilon}}^\top\mathbf{X}^{+\top}\mathbf{X}^+\tilde{\boldsymbol{\varepsilon}} \le c_6\log(n)\mathrm{tr}(\mathbf{X}^{+\top}\mathbf{X}^+)\right\},$$

for positive constants c_1 to c_6.

Proof We have $\mathbb{P}(\mathcal{E}^c) \le \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + \mathbb{P}(\mathcal{E}_3^c)$. The bounds $\mathbb{P}(\mathcal{E}_2^c) \le e^{-cn}$ and $\mathbb{P}(\mathcal{E}_3^c) \le e^{-cn}$ follow immediately from Lemma 23 in Appendix C.1 above, using the fact that $\tilde{\boldsymbol{\varepsilon}}$ has independent entries with sub-Gaussian constants bounded by an absolute constant. Considering $\mathbb{P}(\mathcal{E}_1^c)$, we have

$$\mathbb{P}(\mathcal{E}_1^c) \le \mathbb{P}\{\sigma_n^2(\mathbf{X}) \le c_1\mathrm{tr}(\Sigma_E)\} + \mathbb{P}\{\|\mathbf{E}\|^2 \ge c_2\mathrm{tr}(\Sigma_E)\} + \mathbb{P}\{c_3 n \le \sigma_K^2(\tilde{\mathbf{Z}}) \le \|\tilde{\mathbf{Z}}\|^2 \le c_4 n\}$$

The three terms above can be bounded as follows. Recall that we assume $n > CK$ and $\mathrm{r_e}(\Sigma_E) > Cn$ for some $C > 1$ large enough.

1. Since $\mathrm{r_e}(\Sigma_E) > Cn$, Proposition 14 can be applied to conclude
$$\mathbb{P}\{\sigma_n^2(\mathbf{X}) \le c_1 \mathrm{tr}(\Sigma_E)\} \le 2e^{-cn}.$$

2. By Assumption 3, $\mathbf{E} = \tilde{\mathbf{E}}\Sigma_E^{1/2}$, where $\tilde{\mathbf{E}}$ has independent entries with zero mean, unit variance, and sub-Gaussian constants bounded by an absolute constant. Thus,
$$\|\mathbf{E}\|^2 = \|\mathbf{E}\mathbf{E}^\top\| = \|\tilde{\mathbf{E}}\Sigma_E\tilde{\mathbf{E}}^\top\|,$$
and by applying Theorem 21 with $\tilde{\mathbf{E}}$ and Σ_E we find that with probability at least $1 - 2e^{-cn}$,
$$\|\mathbf{E}\|^2 \le \mathrm{tr}(\Sigma_E) + c'\|\Sigma_E\| n = \mathrm{tr}(\Sigma_E) \cdot (1 + c'n/\mathrm{r_e}(\Sigma_E)) \lesssim \mathrm{tr}(\Sigma_E),$$
where the last inequality holds since $n/\mathrm{r_e}(\Sigma_E) < 1/C$. Thus for $c_2 > 0$,
$$\mathbb{P}\{\|\mathbf{E}\|^2 \ge c_2 \mathrm{tr}(\Sigma_E)\} \le 2e^{-cn}.$$

3. By (78) we have that with probability at least $1 - 2/n$,
$$c_3 n \le \sigma_K^2(\tilde{\mathbf{Z}}) \le \|\tilde{\mathbf{Z}}\|^2 \le c_4 n.$$

Combining the previous three steps shows that $\mathbb{P}(\mathcal{E}_1^c) \le c/n$. ■

C.2.3 Proof of Proposition 14

We will work on the event
$$\mathcal{F} := \{\sigma_n^2(\mathbf{E}U_{(K+1):p}) \ge c_4 \mathrm{tr}(\Sigma_E),\ \|\tilde{\mathbf{Z}}\|^2 \le c_5 n\},$$
where $U_{(K+1):p} \in \mathbb{R}^{p\times(p-K)}$ has columns equal to the orthonormal eigenvectors of Σ_X corresponding to the smallest $p - K$ eigenvalues.

Bounding $\mathbb{P}(\mathcal{F})$*:* By Assumption 3, $\mathbf{E} = \tilde{\mathbf{E}}\Sigma_E^{1/2}$, where $\tilde{\mathbf{E}}$ has independent sub-Gaussian entries with zero mean, unit variance, sub-Gaussian constants bounded by an absolute constant. Thus, letting
$$Q = U_{(K+1):p}U'_{(K+1):p},$$
we have
$$\sigma_n^2(\mathbf{E}U_{(K+1):p}) = \lambda_n(\mathbf{E}Q\mathbf{E}^\top) = \lambda_n(\tilde{\mathbf{E}}\Sigma_E^{1/2}Q\Sigma_E^{1/2}\tilde{\mathbf{E}}^\top).$$
We can now apply Theorem 21, stated and proved above in Section A, with $\tilde{\mathbf{E}}$ and $\Sigma_E^{1/2}Q\Sigma_E^{1/2}$. Noting that $M = \max_{ij}\|\tilde{\mathbf{E}}\|_{\psi_2}$ is bounded by an absolute constant by Assumption 3, this implies that with probability at least $1 - 2e^{-cn}$,
$$\sigma_n^2(\mathbf{E}U_{(K+1):p}) \ge \mathrm{tr}(\Sigma_E^{1/2}Q\Sigma_E^{1/2})/2 - c'\|\Sigma_E^{1/2}Q\Sigma_E^{1/2}\| n. \tag{88}$$

Since Q is a projection matrix, $\|\Sigma_E^{1/2} Q \Sigma_E^{1/2}\| \leq \|\Sigma_E\| \|Q\| = \|\Sigma_E\|$. Furthermore,

$$\begin{aligned}
\operatorname{tr}(\Sigma_E^{1/2} Q \Sigma_E^{1/2}) &= \operatorname{tr}(\Sigma_E Q) \\
&= \operatorname{tr}(\Sigma_E) - \operatorname{tr}(\Sigma_E (I - Q)) \\
&\geq \operatorname{tr}(\Sigma_E) - K \|\Sigma_E (I - Q)\| && (\text{since } \operatorname{rank}(I - Q) = K) \\
&\geq \operatorname{tr}(\Sigma_E) - K \|\Sigma_E\| \|I - Q\| \\
&= \operatorname{tr}(\Sigma_E) - K \|\Sigma_E\| && (\text{since } \|I - Q\| = 1) \\
&\geq \operatorname{tr}(\Sigma_E) - n \|\Sigma_E\|. && (\text{since } n \geq K)
\end{aligned}$$

Plugging these two results into (88), we find that with probability at least $1 - 2e^{-cn}$,

$$\sigma_n^2(\mathbf{E} U_{(K+1):p}) \geq \operatorname{tr}(\Sigma_E)/2 - (1/2 + c') n \|\Sigma_E\| = \operatorname{tr}(\Sigma_E) \cdot [1/2 - (1/2 + c') n / \mathrm{r_e}(\Sigma_E)] \gtrsim \operatorname{tr}(\Sigma_E), \tag{89}$$

where in the last inequality we use that $n/\mathrm{r_e}(\Sigma_E) < 1/C$ and choose C large enough.

Also, since $\tilde{\mathbf{Z}}$ has independent rows with entries that have zero mean, unit variance, and sub-Gaussian constants bounded by an absolute constant, we have that by Theorem 4.6.1 of Vershynin (2019),

$$\|\tilde{\mathbf{Z}}\|^2 \leq c_2 n,$$

with probability at least $1 - e^{-c'n}$. Combining this with 89 we conclude that

$$\mathbb{P}(\mathcal{F}) \geq 1 - c e^{-c'n}.$$

Bounding $\sigma_n(\mathbf{X})$ on $\mathcal{F}$: We now show that $\sigma_n^2(\mathbf{X}) \gtrsim \operatorname{tr}(\Sigma_E)$ holds on the event $\mathcal{F}$. Let $\Sigma_X = U D U^\top$ with $U \in \mathbb{R}^{p \times p}$ orthogonal and $D = \operatorname{diag}(\lambda_1(\Sigma_X), \ldots, \lambda_p(\Sigma_X))$. Define $U_K \in \mathbb{R}^{p \times K}$ to be the sub-matrix of U containing the first K columns, and define $U_{(K+1):p}$ to be composed of the last $p - K$ columns of U. Then

$$I_p = U U^\top = U_K U_K^\top + U_{(K+1):p} U_{(K+1):p}^\top,$$

so

$$\lambda_n(\mathbf{X}\mathbf{X}^\top) = \lambda_n(\mathbf{X} U_K U_K^\top \mathbf{X}^\top + \mathbf{X} U_{(K+1):p} U_{(K+1):p}^\top \mathbf{X}^\top) \geq \lambda_n(\mathbf{X} U_{(K+1):p} U_{(K+1):p}^\top \mathbf{X}^\top),$$

where we use the min-max formula for eigenvalues in the last step. This implies

$$\sigma_n(\mathbf{X}) \geq \sigma_n(\mathbf{X} U_{(K+1):p}). \tag{90}$$

By Weyl's inequality for singular values, and using $\mathbf{X} = \mathbf{Z} A^\top + \mathbf{E}$,

$$|\sigma_n(\mathbf{X} U_{(K+1):p}) - \sigma_n(\mathbf{E} U_{(K+1):p})| \leq \|\mathbf{Z} A^\top U_{(K+1):p}\|,$$

so by (90),

$$\sigma_n(\mathbf{X}) \geq \sigma_n(\mathbf{X} U_{(K+1):p}) \geq \sigma_n(\mathbf{E} U_{(K+1):p}) - \|\mathbf{Z} A^\top U_{(K+1):p}\| \gtrsim \sqrt{\operatorname{tr}(\Sigma_E)} - \|\mathbf{Z} A^\top U_{(K+1):p}\|, \tag{91}$$

where the last inequality holds on the event $\mathcal{F}$. We show below that $\|\mathbf{Z}A^\top U_{(K+1):p}\| \lesssim \sqrt{n\|\Sigma_E\|}$ on $\mathcal{F}$, which implies that

$$\sigma_n(\mathbf{X}) \gtrsim \sqrt{\mathrm{tr}(\Sigma_E)} - c\sqrt{n\|\Sigma_E\|} = \sqrt{\mathrm{tr}(\Sigma_E)} \cdot (1 - c\sqrt{n/\mathrm{r_e}(\Sigma_E)}) \gtrsim \sqrt{\mathrm{tr}(\Sigma_E)},$$

where in the last inequality we use that $n/\mathrm{r_e}(\Sigma_E) < 1/C$ and choose C large enough.

Upper bound of $\|\mathbf{Z}A^\top U_{(K+1):p}\|$*:* On the event $\mathcal{F}$,

$$\|\mathbf{Z}A^\top U_{(K+1):p}\|^2 = \|\tilde{\mathbf{Z}}\Sigma_Z^{1/2}A^\top U_{(K+1):p}\| \le \|\tilde{\mathbf{Z}}\|^2\|\Sigma_Z^{1/2}A^\top U_{(K+1):p}\|^2 \lesssim n\|\Sigma_Z^{1/2}A^\top U_{(K+1):p}\|^2. \tag{92}$$

Furthermore, using $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$, and that $U_{(K+1):p}^\top \Sigma_X U_{(K+1):p} = D_{(K+1):p}$ where we define $D_{(K+1):p} \coloneqq \mathrm{diag}(\lambda_{K+1}(\Sigma_X), \dots, \lambda_p(\Sigma_X))$,

$$\begin{aligned}
\|\Sigma_Z^{1/2}A^\top U_{(K+1):p}\|^2 &= \|U_{(K+1):p}^\top A\Sigma_Z A^\top U_{(K+1):p}\| \\
&= \|U_{(K+1):p}^\top \Sigma_X U_{(K+1):p} - U_{(K+1):p}^\top \Sigma_E U_{(K+1):p}\| \\
&= \|D_{(K+1):p} - U_{(K+1):p}^\top \Sigma_E U_{(K+1):p}\| \\
&\le \lambda_{K+1}(\Sigma_X) + \|U_{(K+1):p}^\top \Sigma_E U_{(K+1):p}\| \\
&\le \lambda_{K+1}(\Sigma_X) + \|\Sigma_E\|\|U_{(K+1):p}^\top U_{(K+1):p}\| \\
&= \lambda_{K+1}(\Sigma_X) + \|\Sigma_E\|,
\end{aligned}$$

where we use $U_{(K+1):p}^\top U_{(K+1):p} = I_{p-K}$ in the last step. Thus, using that

$$\lambda_{K+1}(\Sigma_X) = \lambda_{K+1}(\Sigma_X) - \lambda_{K+1}(A\Sigma_Z A^\top) \le \|\Sigma_E\|$$

by Weyl's inequality and the fact that $\lambda_{K+1}(A\Sigma_Z A^\top) = 0$, we find

$$\|\Sigma_Z^{1/2}A^\top U_{(K+1):p}\|^2 \le 2\|\Sigma_E\|.$$

Combining this with (92), we find that on $\mathcal{F}$,

$$\|\mathbf{Z}A^\top U_{(K+1):p}\| \lesssim \sqrt{n\|\Sigma_E\|}.$$

■

C.3 Proof of Theorem 19 from Section 4.4

Let $D_K = U_K^\top \Sigma_X U_K = \mathrm{diag}(\lambda_1(\Sigma_X), \dots, \lambda_K(\Sigma_X))$ and note that since A and Σ_Z are rank K by Assumption 2,

$$\lambda_K(\Sigma_X) \ge \lambda_K(A\Sigma_Z A^\top) \ge \lambda_K(\Sigma_Z)\lambda_K(AA^\top) > 0,$$

and thus D_K is invertible. Furthermore, define $\eta = y - X^\top\alpha^*$ with variance $\sigma_\eta^2 = \mathbb{E}[\eta^2]$, and the sample version $\boldsymbol{\eta} = \mathbf{y} - \mathbf{X}\alpha^*$. We work on the event $\mathcal{D} \coloneqq \mathcal{D}_1 \cap \mathcal{D}_2$, where

$$\mathcal{D}_1 \coloneqq \left\{\sigma_K^2(\mathbf{X}U_K D_K^{-1/2}) \gtrsim n,\ \|\mathbf{X}\Sigma_X^{-1/2}\|^2 \lesssim p\right\},$$

and

$$\mathcal{D}_2 \coloneqq \left\{ \|(\mathbf{X}U_K D_K^{-1/2})^+ \boldsymbol{\eta}\|^2 \lesssim \log(n) \cdot \sigma_\eta^2 \cdot \mathrm{tr}[(\mathbf{X}U_K D_K^{-1/2})^{+\top} (\mathbf{X}U_K D_K^{-1/2})^+] \right\}.$$

As the last step of this proof, we will show that $\mathbb{P}(\mathcal{D}) \geq 1 - c'/n$.

Letting $\eta \coloneqq y - X^\top \alpha^*$, we have

$$\mathbb{E}[X\eta] = \mathbb{E}[Xy] - \mathbb{E}[XX^\top]\alpha^* = \Sigma_{Xy} - \Sigma_X \Sigma_X^+ \Sigma_{Xy} = 0, \tag{93}$$

where we used (41) in the last step. Thus,

$$\begin{aligned} R(\widetilde{\alpha}_{\mathrm{PCR}}) &\coloneqq \mathbb{E}[(X^\top \widetilde{\alpha}_{\mathrm{PCR}} - y)^2] \\ &= \mathbb{E}\left[(X^\top \widetilde{\alpha}_{\mathrm{PCR}} - X^\top \alpha^* - \eta)^2\right] \\ &= \mathbb{E}\left[(X^\top \widetilde{\alpha}_{\mathrm{PCR}} - X^\top \alpha^*)^2\right] + \mathbb{E}[\eta^2] \qquad \text{(by 93)} \\ &= \|\widetilde{\alpha}_{\mathrm{PCR}} - \alpha^*\|_{\Sigma_X}^2 + R(\alpha^*). \end{aligned} \tag{94}$$

Defining the projection matrix $P = U_K U_K^\top$, and writing

$$\mathbf{y} = \mathbf{X}\alpha^* + \boldsymbol{\eta} = \mathbf{X}P\alpha^* + \mathbf{X}(I_p - P)\alpha^* + \boldsymbol{\eta},$$

we find

$$\begin{aligned} \widetilde{\alpha}_{\mathrm{PCR}} &= U_K(\mathbf{X}U_K)^+ \mathbf{y} \\ &= U_K(\mathbf{X}U_K)^+ \mathbf{X}P\alpha^* + U_K(\mathbf{X}U_K)^+ \mathbf{X}(I_p - P)\alpha^* + U_K(\mathbf{X}U_K)^+ \boldsymbol{\eta}. \end{aligned}$$

From the fact that $\mathbf{X}U_K$ is an $n \times K$ matrix with $K < n$ and $\mathrm{rank}(\mathbf{X}U_K) = K$ on the event $\mathcal{D}_1$, we have $(\mathbf{X}U_K)^+ \mathbf{X}U_K = I_K$ by Lemma 32 of Appendix E below. Thus, using $P = U_K U_K^\top$ we have $(\mathbf{X}U_K)^+ \mathbf{X}P = U_K^\top$. Applying this in the previous display, we find

$$\widetilde{\alpha}_{\mathrm{PCR}} = P\alpha^* + U_K(\mathbf{X}U_K)^+ \mathbf{X}(I_p - P)\alpha^* + U_K(\mathbf{X}U_K)^+ \boldsymbol{\eta}.$$

It thus follows from the decomposition (94) that

$$\begin{aligned} R(\widetilde{\alpha}_{\mathrm{PCR}}) - R(\alpha^*) &= \|\widetilde{\alpha}_{\mathrm{PCR}} - \alpha^*\|_{\Sigma_X}^2 \\ &\lesssim \|(I_p - P)\alpha^*\|_{\Sigma_X}^2 + \|U_K(\mathbf{X}U_K)^+ \mathbf{X}(I_p - P)\alpha^*\|_{\Sigma_X}^2 + \|U_K(\mathbf{X}U_K)^+ \boldsymbol{\eta}\|_{\Sigma_X}^2 \\ &=: B_1 + B_2 + V. \end{aligned} \tag{95}$$

Bounding B_1: We find

$$B_1 = \|\Sigma_X^{1/2}(I_p - P)\alpha^*\|^2 \leq \|\Sigma_X^{1/2}(I_p - P)\|^2 \|\alpha^*\|^2 = \|(I - P)\Sigma_X(I - P)\| \|\alpha^*\|^2. \tag{96}$$

Since $I - P$ is a projection onto the span of the last $p - K$ eigenvectors of Σ_X with eigenvalues $\lambda_{K+1}(\Sigma_X), \ldots, \lambda_p(\Sigma_X)$, we have $\|(I - P)\Sigma_X(I - P)\| = \lambda_{K+1}(\Sigma_X)$. By Weyl's inequality,

$$\lambda_{K+1}(\Sigma_X) = \lambda_{K+1}(\Sigma_X) - \lambda_{K+1}(A\Sigma_Z A^\top) \leq \|\Sigma_E\|,$$

where we used that $\lambda_{K+1}(A\Sigma_Z A^\top) = 0$ in the first step since $\text{rank}(A\Sigma_Z A^\top) = K$. Thus

$$\|\Sigma_X^{1/2}(I_p - P)\|^2 \le \|\Sigma_E\|,$$

and combining this with (96) we find

$$B_1 \le \|\Sigma_E\|\|\alpha^*\|^2. \tag{97}$$

Bounding B_2: Recalling $D_K = U_K^\top \Sigma_X U_K$,

$$\begin{aligned} B_2 &= \alpha^{*\top}(I_p - P)\mathbf{X}^\top(\mathbf{X}U_K)^{+\top}U_K^\top\Sigma_X U_K(\mathbf{X}U_K)^+\mathbf{X}(I-P)\alpha^* \\ &= \|D_K^{1/2}(\mathbf{X}U_K)^+\mathbf{X}(I_p - P)\alpha^*\|^2. \end{aligned} \tag{98}$$

Observe that by Lemma 32 of Appendix E,

$$(\mathbf{X}U_K D_K^{-1/2})^+ = [(\mathbf{X}U_K)^+(\mathbf{X}U_K)D_K^{-1/2}]^+ \cdot [\mathbf{X}U_K D_K^{-1/2} D_K^{1/2}]^+ = D_K^{1/2}(\mathbf{X}U_K)^+, \tag{99}$$

where we used that $\mathbf{X}U_K$ is a full rank $n \times K$ matrix with $K < n$ so $(\mathbf{X}U_K)^+(\mathbf{X}U_K) = I_K$. Using this in (98) yields

$$\begin{aligned} B_2 &= \|(\mathbf{X}U_K D_K^{-1/2})^+\mathbf{X}(I_p - P)\alpha^*\|^2 \\ &\le \frac{\|\mathbf{X}(I_p - P)\alpha^*\|^2}{\sigma_K^2(\mathbf{X}U_K D_K^{-1/2})} \\ &\le \frac{\|\mathbf{X}\Sigma_X^{-1/2}\|^2}{\sigma_K^2(\mathbf{X}U_K D_K^{-1/2})} \cdot \|\Sigma_X^{1/2}(I_p - P)\alpha^*\|^2 \\ &\lesssim \frac{p}{n}\|\Sigma_X^{1/2}(I_p - P)\alpha^*\|^2, \end{aligned}$$

where the last step holds on $\mathcal{D}$. Recalling that $\|\Sigma_X^{1/2}(I_p - P)\alpha^*\|^2 = B_1$ and using (97), we find that

$$B_2 \lesssim \|\Sigma_E\| \cdot \|\alpha^*\|^2 \frac{p}{n}. \tag{100}$$

Bounding V: We have on $\mathcal{D}$,

$$\begin{aligned} V &= \boldsymbol{\eta}^\top(\mathbf{X}U_K)^{+\top}U_K^\top\Sigma_X U_K(\mathbf{X}U_K)^+\boldsymbol{\eta} \\ &= \boldsymbol{\eta}^\top(\mathbf{X}U_K)^{+\top}D_K(\mathbf{X}U_K)^+\boldsymbol{\eta} \\ &= \|D_K^{1/2}(\mathbf{X}U_K)^+\boldsymbol{\eta}\|^2 \\ &= \|(\mathbf{X}U_K D_K^{-1/2})^+\boldsymbol{\eta}\|^2 && \text{(by (99))} \\ &\lesssim \sigma_\eta^2 \cdot \log(n) \cdot \text{tr}[(\mathbf{X}U_K D^{-1/2})^{+\top}(\mathbf{X}U_K D^{-1/2})^+] && \text{(on } \mathcal{D}_2\text{)} \\ &\le \sigma_\eta^2 \cdot \log(n) \cdot K \cdot \|(\mathbf{X}U_K D^{-1/2})^+\|^2 && \text{(since rank}(\mathbf{X}U_K D^{-1/2}) = K\text{)} \\ &= \sigma_\eta^2 \cdot \frac{K\log n}{\sigma_K^2(\mathbf{X}U_K D^{-1/2})} \\ &\lesssim \sigma_\eta^2 \cdot \frac{K\log n}{n}. && \text{(on } \mathcal{D}_1\text{)}. \end{aligned}$$

Recalling $\eta = y - X^\top \alpha^*$ so $\sigma_\eta^2 = R(\alpha^*)$,

$$V \lesssim R(\alpha^*) \cdot \frac{K \log n}{n}. \tag{101}$$

Combining this with (97) and (100) proves (33).

In the case $\Sigma_E = 0$, the bound (34) follows immediately from (33). When $\lambda_p(\Sigma_E) > 0$, Lemma 5 of Section 3.2 implies

$$R(\alpha^*) \le \sigma_\varepsilon^2 + \frac{\|\beta\|^2}{\xi}.$$

When $\lambda_p(\Sigma_E) > 0$, we also have that

$$\|\alpha^*\|^2 \le \kappa(\Sigma_E)\beta^\top (A^\top A)^{-1}\beta \le \frac{1}{\lambda_p(\Sigma_E)} \cdot \frac{\|\beta\|_{\Sigma_Z}^2}{\xi}.$$

Plugging the last two displays into (33) gives

$$\begin{aligned} R_{\mathrm{PCR}}(\hat{\beta}) - R(\alpha^*) &\lesssim \kappa(\Sigma_E)\frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \cdot \frac{p}{n} + \frac{\|\beta\|_{\Sigma_Z}^2}{\xi}\frac{K \log n}{n} + \sigma_\varepsilon^2 \frac{K \log n}{n} \\ &\lesssim \kappa(\Sigma_E)\frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \cdot \frac{p}{n} + \sigma_\varepsilon^2 \frac{K \log n}{n}, \end{aligned}$$

where in the second step we use that

$$K \log n < c \cdot n \lesssim p \le \kappa(\Sigma_E) p.$$

This proves (35). All that remains is to bound the probability of the event $\mathcal{D}$.

Bounding $\mathbb{P}(\mathcal{D})$*:* We first bound the probability $\mathbb{P}(\mathcal{D}_1)$. Note that the matrix $\mathbf{X} U_K D_K^{-1/2}$ has independent Gaussian rows $D_K^{-1/2} U_K^\top X_i$, with covariance

$$\mathbb{E}[D_K^{-1/2} U_K^\top X_i X_i^\top U_K D_K^{-1/2}] = D_K^{-1/2} U_K^\top \Sigma_X U_K D_K^{-1/2} = D_K^{-1/2} D_K D_K^{-1/2} = I_K,$$

and so $\mathbf{X} U_K D_K^{-1/2}$ i.i.d. $N(0,1)$ entries. Thus, by Theorem 4.6.1 of Vershynin (2019), with probability at least $1 - 2/n$,

$$\sigma_K(\mathbf{X} U_K D_K^{-1/2}) \ge \sqrt{n} - c(\sqrt{K} + \sqrt{\log n}) = \sqrt{n} \cdot [1 - c\sqrt{K/n} - c\sqrt{\log(n)/n}] \gtrsim \sqrt{n}, \tag{102}$$

where in the last step we use the assumption that $n > CK > C$ and choose C large enough.

Similarly, $\mathbf{X}\Sigma_X^{-1/2}$ is a $n \times p$ matrix with i.i.d. $N(0,1)$ entries, so again by by Theorem 4.6.1 of Vershynin (2019), with probability at least $1 - 2e^{-n}$,

$$\|\mathbf{X}\Sigma_X^{-1/2}\| \le \sqrt{n} + c(\sqrt{p} + \sqrt{n}) \lesssim \sqrt{p}. \tag{103}$$

Using a union bound to combine this with (102), we find

$$\mathbb{P}(\mathcal{D}_1) \ge 1 - c'/n,$$

for some $c' > 0$.

To bound $\mathbb{P}(\mathcal{D}_2)$, first note that by (93) and the assumption that (X, y) are Gaussian, $\mathbf{X}$ and $\boldsymbol{\eta}$ are independent. Furthermore, $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}/\sigma_\eta$ has independent $N(0,1)$ entries. We can thus apply Lemma 23 from Appendix C.1 above with

$$M = (\mathbf{X}U_K D_K^{-1/2})^{+\top}(\mathbf{X}U_K D_K^{-1/2})^+$$

to conclude that with probability at least $1 - e^{-cn}$,

$$\|(\mathbf{X}U_K D_K^{-1/2})^+\boldsymbol{\eta}\|^2 = \boldsymbol{\eta}^\top M \boldsymbol{\eta} = \sigma_\eta^2 \tilde{\boldsymbol{\eta}}^\top M \tilde{\boldsymbol{\eta}} \lesssim \sigma_\eta^2 \cdot \log(n) \cdot \operatorname{tr}(M),$$

and so $\mathbb{P}(\mathcal{D}_2^c) \le e^{-cn}$. ■

C.4 Detailed Comparison of the Bias and Variance Terms in Section 4.3

In this sections we give a detailed comparison between our Theorem 16 and Theorem 4 in Bartlett et al. (2020). We assume throughout this section that the matrices Σ_X and Σ_E are invertible and the condition number $\kappa(\Sigma_E)$ of the matrix Σ_E is bounded above by an absolute constant c_1.

First define the effective ranks

$$r_k(\Sigma_X) \coloneqq \frac{\sum_{i>k}\lambda_i(\Sigma_X)}{\lambda_{i+1}(\Sigma_X)}, \qquad R_k(\Sigma_X) \coloneqq \frac{\left(\sum_{i>k}\lambda_i(\Sigma_X)\right)^2}{\sum_{i>k}\lambda_i^2(\Sigma_X)}.$$

The bound of Bartlett et al. (2020) is stated to hold for probability at least $1-\delta$ for a general $\delta < 1$ such that $\log(1/\delta) > n/c$ for an absolute constant $c > 1$. Taking $\delta = e^{-c'n}$ (for an appropriate c') to ease comparison with our results, the bound then states that with when model (5) holds, (X, y) are jointly Gaussian, $\operatorname{rank}(\Sigma_X) \ge n$, and n is large enough, with probability at least $1 - e^{-c'n}$,

$$R(\widehat{\alpha}) - R(\alpha^*) \lesssim B + V,$$

where

$$B \coloneqq \|\alpha^*\|^2 \|\Sigma_X\| \max\left\{\sqrt{\frac{r_0(\Sigma_X)}{n}}, \frac{r_0(\Sigma_X)}{n}, 1\right\}, \tag{104}$$

and

$$V \coloneqq \sigma_\varepsilon^2 \log(n)\left(\frac{n}{R_{K^*}(\Sigma_X)} + \frac{K^*}{n}\right) \tag{105}$$

are bounds on the bias and variance respectively, and

$$K^* = \min\{k \ge 0 : r_k(\Sigma_X)/n \ge b\}, \tag{106}$$

where $b > 1$ is an absolute constant.

We now compare these two terms to the corresponding terms in our bound in Theorem 16.

C.4.1 Comparison of Variance Terms

We first compare the variance term V to corresponding variance term in our Theorem 16, display (27). Note that as long as the SNR

$$\xi \coloneqq \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\|$$

grows fast enough, $K^* = K$ for large enough n, where K is the dimension of the latent variables $Z \in \mathbb{R}^K$ in the factor regression model.

Lemma 25. *If $K/n = o(1)$, $\mathrm{r_e}(\Sigma_E)/n \to \infty$, and $\xi \to \infty$, such that $\xi^{-1}\mathrm{r_e}(\Sigma_E)/n = o(1)$, then $K^* = K$ for all n large enough.*

Thus, under the conditions stated in Lemma 25 and for n large enough,

$$V \coloneqq \sigma_\varepsilon^2 \log(n)\left(\frac{n}{R_K(\Sigma_X)} + \frac{K}{n}\right).$$

Using the convexity of $x \mapsto x^2$, we can bound $R_K(\Sigma_X)$ above via

$$R_K(\Sigma_X) = \frac{\left(\sum_{i=K+1}^p \lambda_i(\Sigma_X)\right)^2}{\sum_{i=K+1}^p \lambda_i^2(\Sigma_X)} \le \frac{(p-K)\sum_{i=K+1}^p \lambda_i^2(\Sigma_X)}{\sum_{i=K+1}^p \lambda_i^2(\Sigma_X)} \le p.$$

Thus,

$$V \ge \sigma_\varepsilon^2 \log(n)\left(\frac{n}{p} + \frac{K}{n}\right). \tag{107}$$

When $\kappa(\Sigma_E) < c_1$, $p \lesssim \mathrm{r_e}(\Sigma_E) \le p$, and so the variance term in the bound of our Theorem 16 is

$$\sigma_\varepsilon^2 \log(n)\left(\frac{n}{r_0(\Sigma_E)} + \frac{K}{n}\right) \lesssim \sigma_\varepsilon^2 \log(n)\left(\frac{n}{p} + \frac{K}{n}\right).$$

Thus, comparing with (107), we see that under the stated conditions our variance bound is the same as that of Bartlett et al. (2020), up to absolute constants.

Proof [Proof of Lemma 25] We will prove that

$$\frac{r_\ell(\Sigma_X)}{n} \le \frac{K}{n}(1+\xi^{-1}) + \frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}, \qquad \text{for } 0 \le \ell \le K-1 \tag{108}$$

and that

$$\frac{r_K(\Sigma_X)}{n} \ge \frac{\mathrm{r_e}(\Sigma_E)}{n} - \frac{K}{n}. \tag{109}$$

Together with the definition of K^* in (106), these two bounds imply Lemma 25.

First note that for $0 \le \ell \le K$,

$$\begin{aligned}
\sum_{i=\ell+1}^p \lambda_i(\Sigma_X) &= \mathrm{tr}(\Sigma_X) - \sum_{i=1}^\ell \lambda_i(\Sigma_X) \\
&= \mathrm{tr}(\Sigma_E) + \mathrm{tr}(A\Sigma_Z A^\top) - \sum_{i=1}^\ell \lambda_i(\Sigma_X) \\
&= \mathrm{tr}(\Sigma_E) + \sum_{i=\ell+1}^K \lambda_i(A\Sigma_Z A^\top) + \sum_{i=1}^\ell (\lambda_i(A\Sigma_Z A^\top) - \lambda_i(\Sigma_X)),
\end{aligned} \tag{110}$$

where the sums from $\ell+1$ to K and from 1 to ℓ are defined to be zero when $\ell = K$ and $\ell = 0$, respectively.

Proof of (108): By Weyl's inequality,

$$|\lambda_i(A\Sigma_Z A^\top) - \lambda_i(\Sigma_X)| \le \|\Sigma_E\|, \tag{111}$$

so by (110),

$$\begin{aligned}\sum_{i=\ell+1}^{p} \lambda_i(\Sigma_X) &\le \operatorname{tr}(\Sigma_E) + (K-\ell)\lambda_{\ell+1}(A\Sigma_Z A^\top) + \ell\|\Sigma_E\| \\ &\le \operatorname{tr}(\Sigma_E) + K\lambda_{\ell+1}(A\Sigma_Z A^\top) + K\|\Sigma_E\|. \end{aligned}\tag{112}$$

From the min-max formula for eigenvalues we have

$$\lambda_{\ell+1}(\Sigma_X) = \min_{S:\dim(S)=\ell+1} \max_{x\in S:\|x\|=1} x^\top \Sigma_X x,$$

where the minimum is taken over all linear subspaces $S \subset \mathbb{R}^p$ with dimension $\ell+1$. Since $x^\top \Sigma_X x \ge x^\top A\Sigma_Z A^\top x$ for any $x \in \mathbb{R}^p$, this implies

$$\lambda_{\ell+1}(\Sigma_X) \ge \lambda_{\ell+1}(A\Sigma_Z A^\top). \tag{113}$$

Combining (112) and (113), we find

$$\begin{aligned} r_\ell(\Sigma_X) &= \frac{\sum_{i=\ell+1}^{p} \lambda_i(\Sigma_X)}{\lambda_{\ell+1}(\Sigma_X)} \\ &\le K\left(1 + \frac{\|\Sigma_E\|}{\lambda_{\ell+1}(A\Sigma_Z A^\top)}\right) + \frac{\operatorname{tr}(\Sigma_E)}{\lambda_{\ell+1}(A\Sigma_Z A^\top)} \\ &\le K\left(1 + \frac{\|\Sigma_E\|}{\lambda_K(A\Sigma_Z A^\top)}\right) + \frac{\operatorname{tr}(\Sigma_E)}{\lambda_K(A\Sigma_Z A^\top)} \\ &= K(1+\xi^{-1}) + \xi^{-1}\mathrm{r_e}(\Sigma_E), \end{aligned}$$

which completes the proof of (108).

Proof of (109): Equation (110) for $\ell = K$ is

$$\sum_{i=K+1}^{p} \lambda_i(\Sigma_X) = \operatorname{tr}(\Sigma_E) + \sum_{i=1}^{K} (\lambda_i(A\Sigma_Z A^\top) - \lambda_i(\Sigma_X)).$$

Again using (111),

$$\sum_{i=K+1}^{p} \lambda_i(\Sigma_X) \ge \operatorname{tr}(\Sigma_E) - K\|\Sigma_E\|. \tag{114}$$

Since

$$\begin{aligned} \lambda_{K+1}(\Sigma_X) &= \lambda_{K+1}(\Sigma_X) - \lambda_{K+1}(A\Sigma_Z A^\top) \qquad (\text{since } \lambda_{K+1}(A\Sigma_Z A^\top) = 0) \\ &\le \|\Sigma_E\| \qquad (\text{Weyl's inequality}). \end{aligned}\tag{115}$$

Combining (114) and (115), we find

$$r_K(\Sigma_X) = \frac{\sum_{i=K+1}^{p} \lambda_i(\Sigma_X)}{\lambda_{K+1}(\Sigma_X)} \geq \mathrm{r_e}(\Sigma_E) - K,$$

which proves (109).

■

C.4.2 Comparison of Bias Terms

A more interesting comparison arises between the bias term B and the corresponding bias term in Theorem 16, display (27). Here we will see how the approach we take in this paper, explicitly taking advantage of the structure of the factor regression model, leads to a stronger bound under certain conditions

Lemma 26. *Suppose $\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| > 1$ and A, Σ_Z, Σ_E are all full rank. Then*

$$B \geq \left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \|\beta\|^2_{\Sigma_Z} \max\left(\sqrt{\frac{r_0(\Sigma_X)}{n}}, \frac{r_0(\Sigma_X)}{n}\right), \tag{116}$$

where

$$\frac{r_0(\Sigma_X)}{n} \geq \frac{1}{2}\frac{r_0(A\Sigma_Z A^\top)}{n} + \frac{1}{2\kappa(A\Sigma_Z A^\top)}\frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}. \tag{117}$$

In particular, if $\xi > c_1 > 1$ and $\kappa(\Sigma_E) < c_2$, $\kappa(A\Sigma_Z A^\top) < c_2$ for absolute constants c_1, c_2,

$$B \gtrsim \|\beta\|^2_{\Sigma_Z} \max\left(\sqrt{\frac{1}{\xi}\frac{p}{n}}, \frac{1}{\xi}\frac{p}{n}\right). \tag{118}$$

Compared to our bias bound $\|\beta\|^2_{\Sigma_Z} p/(n \cdot \xi)$ in Theorem 16, there is an additional quantity $r_0(A\Sigma_Z A^\top)/n$ of order $O(K/n)$. Ignoring this quantity, provided both $\kappa(\Sigma_E)$ and $\kappa(A\Sigma_Z A^\top)$ are uniformly bounded, we obtain the lower bound (118). When $p/(n \cdot \xi) < 1$, this rate is worse by a factor $\sqrt{p/(n \cdot \xi)}$, compared to the bias term $\|\beta\|^2_{\Sigma_Z} p/(n \cdot \xi)$ in Theorem 16.

Proof [Proof of Lemma 26] Using that A, Σ_Z, Σ_E are all full rank, by (64) above,

$$\|\alpha^*\|^2 \geq \left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \cdot \beta^\top (A^\top A)^{-1} \beta \geq \left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \frac{\|\beta\|^2_{\Sigma_Z}}{\|A\Sigma_Z A^\top\|}.$$

Thus, using $\|\Sigma_X\| = \|A\Sigma_Z A^\top + \Sigma_E\| \geq \|A\Sigma_Z A^\top\|$,

$$\|\Sigma_X\| \|\alpha^*\|^2 \geq \left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \|\beta\|^2_{\Sigma_Z},$$

which implies (116).

To prove (117), we first recall that $r_0(\Sigma_X) = \mathrm{tr}(\Sigma_X)/\|\Sigma_X\|$ and $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$, which implies that

$$\frac{r_0(\Sigma_X)}{n} = \frac{\mathrm{tr}(A\Sigma_Z A^\top)}{n\|\Sigma_X\|} + \frac{\mathrm{tr}(\Sigma_E)}{n\|\Sigma_X\|}.$$

Observing that $\|\Sigma_X\| \le \|A\Sigma_Z A^\top\| + \|\Sigma_E\| \le 2\|A\Sigma_Z A^\top\|$, where we use that $\|\Sigma_E\| \le \|A\Sigma_Z A^\top\|$ by the assumption $\xi > 1$, we find

$$\begin{aligned}
\frac{r_0(\Sigma_X)}{n} &\ge \frac{1}{2}\frac{r_0(A\Sigma_Z A^\top)}{n} + \frac{1}{2}\frac{\mathrm{tr}(\Sigma_E)}{n\|A\Sigma_Z A^\top\|} \\
&= \frac{1}{2}\frac{r_0(A\Sigma_Z A^\top)}{n} + \frac{1}{2}\frac{\lambda_K(A\Sigma_Z A^\top)}{\|A\Sigma_Z A^\top\|}\frac{\|\Sigma_E\|}{\lambda_K(A\Sigma_Z A^\top)}\frac{\mathrm{tr}(\Sigma_E)}{n\|\Sigma_E\|} \\
&= \frac{1}{2}\frac{r_0(A\Sigma_Z A^\top)}{n} + \frac{1}{2\kappa(A\Sigma_Z A^\top)}\frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n},
\end{aligned}$$

which proves (117). ■

Appendix D. Supplementary Results

D.1 Closed Form Solutions of Min-Norm Estimator and Minimizer of $R(\alpha)$

Lemma 27. *For zero mean random variables $X \in \mathbb{R}^p$ and $y \in \mathbb{R}$, suppose $\Sigma_X := \mathbb{E}[XX^\top]$ and $\sigma_y^2 := \mathbb{E}[y^2]$ are finite, and let $\Sigma_{Xy} = \mathbb{E}[Xy]$. Then $\alpha^* := \Sigma_X^+\Sigma_{Xy}$ is a minimizer of $R(\alpha)$:*

$$R(\alpha^*) = \min_{\alpha\in\mathbb{R}^p} R(\alpha).$$

Proof We have

$$R(\alpha) = \mathbb{E}[(X^\top\alpha - y)^2] = \alpha^\top\Sigma_X\alpha + \sigma_y^2 - 2\alpha^\top\Sigma_{Xy},$$

so since $R(\alpha)$ is convex, α is a minimizer if and only if

$$\nabla_\alpha R(\alpha) = 2\Sigma_X\alpha - 2\Sigma_{Xy} = 0.$$

By (41), $\Sigma_X\alpha^* = \Sigma_{Xy}$, so the claim is proved. ■

For $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \in \mathbb{R}^n$, let

$$\widehat{\alpha} := \arg\min\left\{\|\alpha\| :\ \|\mathbf{X}\alpha - \mathbf{y}\| = \min_u \|\mathbf{X}u - \mathbf{y}\|\right\}.$$

We then have the following result.

Lemma 28. $\widehat{\alpha} = \mathbf{X}^+\mathbf{y}$.

Proof We establish the proof in two steps.

Step 1: Existence and uniqueness of $\widehat{\alpha}$. Since

$$\nabla_u\|\mathbf{X}u - y\|^2 = 2\mathbf{X}^\top\mathbf{X}u - 2\mathbf{X}^\top y,$$

and $\|\mathbf{X}u - y\|^2$ is convex in u, u is a minimizer of $u \mapsto \|\mathbf{X}u - y\|^2$ if and only if

$$\mathbf{X}^\top \mathbf{X}u = \mathbf{X}^\top \mathbf{y}. \tag{119}$$

By the properties of the pseudo-inverse, $\mathbf{X}^\top \mathbf{X}\mathbf{X}^+ = \mathbf{X}^\top$, so

$$\mathbf{X}^\top \mathbf{X}(\mathbf{X}^+ \mathbf{y}) = \mathbf{X}^\top \mathbf{y},$$

and thus $\mathbf{X}^+\mathbf{y}$ is a minimizer of $\|\mathbf{X}u - \mathbf{y}\|$. The set of vectors u satisfying $\mathbf{X}^\top \mathbf{X}u = \mathbf{X}^\top \mathbf{y}$ is also convex, so $\widehat{\alpha}$ is a minimizer of a strictly convex function $\|\cdot\|$ over a non-empty convex set. Such a minimizer exists and is unique, so $\widehat{\alpha}$ exists and is unique.

Step 2: formula for $\widehat{\alpha}$. Since $\widehat{\alpha}$ is a minimizer of $\|\mathbf{X}u - \mathbf{y}\|$, it must satisfy 119, i.e.

$$\mathbf{X}^\top \mathbf{X}\widehat{\alpha} = \mathbf{X}^\top y. \tag{120}$$

We can write

$$\widehat{\alpha} = \mathbf{X}^+\mathbf{X}\widehat{\alpha} + (I - \mathbf{X}^+\mathbf{X})\widehat{\alpha},$$

and using $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$ as well as the fact that $\mathbf{X}^+\mathbf{X}$ is symmetric (see Appendix E), a quick calculation gives

$$\|\widehat{\alpha}\|^2 = \|\mathbf{X}^+\mathbf{X}\widehat{\alpha}\|^2 + \|(I - \mathbf{X}^+\mathbf{X})\widehat{\alpha}\|^2.$$

Thus $\|\mathbf{X}^+\mathbf{X}\widehat{\alpha}\| \leq \|\widehat{\alpha}\|^2$, and also

$$\mathbf{X}^\top \mathbf{X}(\mathbf{X}^+\mathbf{X}\widehat{\alpha}) = \mathbf{X}^\top \mathbf{X}\widehat{\alpha} = \mathbf{X}^\top \mathbf{y},$$

where we used $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$ in the first step and 120 in the second step. Thus $\mathbf{X}^+\mathbf{X}\widehat{\alpha}$ is a minimizer of $\|\cdot\|$ among minimizers of $\|\mathbf{X}u - \mathbf{y}\|$. Since by Step 1 above $\widehat{\alpha}$ is the unique such minimizer, $\mathbf{X}^+\mathbf{X}\widehat{\alpha} = \widehat{\alpha}$. Thus,

$$\begin{aligned}
\widehat{\alpha} &= \mathbf{X}^+\mathbf{X}\widehat{\alpha} \\
&= (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{X}\widehat{\alpha} && \text{(since } \mathbf{X}^+ = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top\text{)} \\
&= (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{y} && \text{(by 120)} \\
&= \mathbf{X}^+\mathbf{y}. && \text{(since } \mathbf{X}^+ = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top\text{)}
\end{aligned}$$

■

D.2 Proof that (5) is a Special Case of (21) in the Gaussian Case

Lemma 29. *Suppose that (X, y) follows model (5) with mean zero and is furthermore jointly Gaussian. Then model (21) holds with $\theta = \alpha^*$ and and error $\eta := y - X^\top \alpha^*$, independent of X, where $\alpha^* = \Sigma_X^+ \Sigma_{Xy}$ is the best linear predictor under model (5).*

Proof We first compute

$$\mathbb{E}[X\eta] = \mathbb{E}[X(y - X^\top \alpha^*)^2] = \mathbb{E}[XX^\top]\alpha^* - \mathbb{E}[Xy] = \Sigma_X \alpha^* - \Sigma_{Xy},$$

where we use that X and y are mean zero in the final step. Using the fact that $\Sigma_X \alpha^* = \Sigma_{Xy}$ from (41) above, we find $\mathbb{E}[X\eta] = 0$ so X and η are uncorrelated, where we again use that (X, y) are mean zero, so η is mean zero. Since X and y are jointly normal, it follows that X and η are jointly normal. Thus, X and η are independent and so model (21) holds as claimed. ■

D.3 Risk of $\widehat{\alpha}$ Under the Factor Regression Model for $p \ll n$

For completeness, we provide a risk bound for the minimum-norm estimator $\widehat{\alpha}$ under the factor regression model in the low-dimensional regime $p \ll n$.

Theorem 30. *Under model 5, suppose that Assumptions 1, 2 & 3 hold. Then if $n > C \cdot p$ for some $C > 0$ large enough and $p \geq K$, with probability at least $1 - c/n$,*

$$R(\widehat{\alpha}) - \sigma_\varepsilon^2 \lesssim \kappa(\Sigma_E) \frac{\|\beta\|_{\Sigma_Z}^2}{\xi} + \frac{p}{n} \sigma_\varepsilon^2 \log n,$$

where $\kappa(\Sigma_E) = \lambda_1(\Sigma_E)/\lambda_p(\Sigma_E)$ is the condition number of Σ_E.

Proof As in the proof of Theorem 16 found in section C.2.2 above,

$$R(\widehat{\alpha}) \leq 2(B_1 + B_2) + 2(V_1 + V_2),$$

where

$$\begin{aligned} B_1 &= \|\Sigma_E^{1/2} \mathbf{X}^+ \mathbf{Z} \beta\|^2 \\ B_2 &= \|\Sigma_Z^{1/2} (A^\top \mathbf{X}^+ \mathbf{Z} - I_K) \beta\|^2 \\ V_1 &= \|\Sigma_E^{1/2} \mathbf{X}^+ \boldsymbol{\varepsilon}\|^2 \\ V_2 &= \|\Sigma_Z^{1/2} A^\top \mathbf{X}^+ \boldsymbol{\varepsilon}\|^2. \end{aligned}$$

We will bound these four terms on the event $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2$, where

$$\mathcal{B}_1 := \{\|\tilde{\mathbf{E}}\|^2 < c_1 n,\ \sigma_K^2(\tilde{\mathbf{Z}}) > c_2 n,\ \sigma_p^2(\tilde{\mathbf{X}}) \geq c_3 n\}$$

and

$$\mathcal{B}_2 := \left\{ \tilde{\boldsymbol{\varepsilon}}^\top \mathbf{X}^{+\top} \Sigma_X \mathbf{X}^+ \tilde{\boldsymbol{\varepsilon}} \leq c_5 \log(n) \cdot \operatorname{tr}(\mathbf{X}^{+\top} \Sigma_X \mathbf{X}^+) \right\}.$$

As the last step of the proof, we will show that $\mathbb{P}(\mathcal{B}) \geq 1 - c/n$.

Bounding the bias component: First observe that since $K < n$, when $\mathbf{Z}$ is full rank, $\mathbf{Z}^+ \mathbf{Z} = I_K$ and so

$$A^\top \mathbf{X}^+ = \mathbf{Z}^+ \mathbf{Z} A^\top \mathbf{X}^+ = \mathbf{Z}^+ (\mathbf{X} - \mathbf{E}) \mathbf{X}^+ = \mathbf{Z}^+ \mathbf{X} \mathbf{X}^+ - \mathbf{Z}^+ \mathbf{E} \mathbf{X}^+.$$

Thus,

$$\begin{aligned} B_2 &= \|(A^\top \mathbf{X}^+ \mathbf{Z} - I_K)\beta\|^2 \\ &= \|(\mathbf{Z}^+ \mathbf{X} \mathbf{X}^+ \mathbf{Z} - I_K)\beta - \mathbf{Z}^+ \mathbf{E} \mathbf{X}^+ \mathbf{Z} \beta\|_{\Sigma_Z}^2 \\ &\leq 2\|(\mathbf{Z}^+ \mathbf{X} \mathbf{X}^+ \mathbf{Z} - I_K)\beta\|_{\Sigma_Z}^2 + 2\|\mathbf{Z}^+ \mathbf{E} \mathbf{X}^+ \mathbf{Z} \beta\|_{\Sigma_Z}^2. \end{aligned} \tag{121}$$

Note that since $p \geq K$, by Assumption 2, $\text{rank}(A) = K$ so by Lemma 32 of Appendix E,

$$A^\top A^{+\top} = I_K. \tag{122}$$

We thus have

$$\begin{aligned}
\|(\mathbf{Z}^+\mathbf{X}\mathbf{X}^+\mathbf{Z} - I_K)\beta\|_{\Sigma_Z}^2 &= \|(\mathbf{Z}^+\mathbf{X}\mathbf{X}^+\mathbf{Z} - \mathbf{Z}^+\mathbf{Z})\beta\|_{\Sigma_Z}^2 \\
&= \|\tilde{\mathbf{Z}}^+(\mathbf{X}\mathbf{X}^+ - I_p)\mathbf{Z}\beta\|^2 \\
&\leq \frac{\|(\mathbf{X}\mathbf{X}^+ - I_p)\mathbf{Z}\beta\|^2}{\sigma_K^2(\tilde{\mathbf{Z}})} \\
&\lesssim \frac{1}{n}\|(\mathbf{X}\mathbf{X}^+ - I_p)\mathbf{Z}\beta\|^2 && \text{(on } \mathcal{B}\text{)} \\
&= \frac{1}{n}\|(\mathbf{X}\mathbf{X}^+ - I_p)\mathbf{Z}A^\top A^{+\top}\beta\|^2 && \text{(by (122))} \\
&= \frac{1}{n}\|(\mathbf{X}\mathbf{X}^+ - I_p)(\mathbf{X} - \mathbf{E})A^{+\top}\beta\|^2 && \text{(since } \mathbf{X} = \mathbf{Z}A^\top + \mathbf{E}\text{)} \\
&= \frac{1}{n}\|(\mathbf{X}\mathbf{X}^+ - I_p)\mathbf{E}A^{+\top}\beta\|^2 && \text{(since } \mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}\text{)} \\
&\leq \frac{1}{n}\|\mathbf{X}\mathbf{X}^+ - I_p\| \cdot \|\mathbf{E}A^{+\top}\beta\|^2 \\
&\leq \frac{1}{n}\|\mathbf{E}A^{+\top}\beta\|^2 \\
&\lesssim \frac{n\|\Sigma_E\|}{n}\frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)} && \text{(on } \mathcal{B} \text{ and by (83))} \\
&= \frac{\|\beta\|_{\Sigma_Z}^2}{\xi}, && \text{(123)}
\end{aligned}$$

where in the penultimate step we used

$$\|A^{+\top}\beta\|^2 \leq \frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)} \tag{124}$$

from (83). We can bound the second term in 121 as follows:

$$\begin{aligned}
\|\mathbf{Z}^+\mathbf{E}\mathbf{X}^+\mathbf{Z}\beta\|_{\Sigma_Z}^2 &= \|\tilde{\mathbf{Z}}^+\mathbf{E}\mathbf{X}^+\mathbf{Z}\beta\|^2 \\
&\leq \frac{\|\mathbf{E}\|^2}{\sigma_K^2(\tilde{\mathbf{Z}})}\|\mathbf{X}^+\mathbf{Z}\beta\|^2 \\
&\lesssim \|\Sigma_E\| \cdot \|\mathbf{X}^+\mathbf{Z}\beta\|^2 && \text{(on } \mathcal{B}\text{)} \\
&= \|\Sigma_E\| \cdot \|\mathbf{X}^+\mathbf{Z}A^\top A^{+\top}\beta\|^2 && \text{(since } A^\top A^{+\top} = I_K\text{)} \\
&= \|\Sigma_E\| \cdot \|\mathbf{X}^+(\mathbf{X} - \mathbf{E})A^{+\top}\beta\|^2 && \text{(since } \mathbf{X} = \mathbf{Z}A^\top + \mathbf{E}\text{)} \\
&\leq 2\|\Sigma_E\| \cdot \|\mathbf{X}^+\mathbf{X}A^{+\top}\beta\|^2 + 2\|\Sigma_E\| \cdot \|\mathbf{X}^+\mathbf{E}A^{+\top}\beta\|^2 \\
&\lesssim \|\Sigma_E\|\|A^{+\top}\beta\|^2 + \|\Sigma_E\|\frac{\|\mathbf{E}\|}{\sigma_p^2(\mathbf{X})}\|A^{+\top}\beta\|^2 && \text{(since } \|\mathbf{X}^+\mathbf{X}\| \leq 1\text{)} \\
&\lesssim \|\Sigma_E\| \cdot \kappa(\Sigma_E)\|A^{+\top}\beta\|^2 \\
&\leq \kappa(\Sigma_E)\frac{\|\beta\|_{\Sigma_Z}^2}{\xi}. && \text{(by (124))}
\end{aligned}$$

Using this and (123) in (121), and using the fact that $\kappa(\Sigma_E) > 1$, we find that on the event $\mathcal{B}$,

$$B_2 \lesssim \kappa(\Sigma_E)\frac{\|\beta\|_{\Sigma_Z}^2}{\xi}. \tag{125}$$

Bounding the variance component: We have

$$\begin{aligned} V_1 + V_2 &= \boldsymbol{\varepsilon}^\top \mathbf{X}^{+\top}\Sigma_X \mathbf{X}^+ \boldsymbol{\varepsilon} \\ &= \sigma_\varepsilon^2 \tilde{\boldsymbol{\varepsilon}}^\top \mathbf{X}^{+\top}\Sigma_X \mathbf{X}^+ \tilde{\boldsymbol{\varepsilon}} && \text{(by Assumption 3)} \\ &\lesssim \sigma_\varepsilon^2 \log(n)\mathrm{tr}(\mathbf{X}^{+\top}\Sigma_X \mathbf{X}^+) && \text{(on } \mathcal{B}_2\text{)} \\ &\le \sigma_\varepsilon^2 \log(n)\cdot p\|\mathbf{X}^{+\top}\Sigma_X \mathbf{X}^+\| && \text{(since } \mathrm{rank}(\mathbf{X}^+) = p\text{)} \\ &= \sigma_\varepsilon^2 \log(n)\cdot p\|\Sigma_X^{1/2}\mathbf{X}^+\|^2. \end{aligned} \tag{126}$$

From Assumption 1, $\mathbf{X} = \tilde{\mathbf{X}}\Sigma_X^{1/2}$, and from Lemma 32 of Appendix E below,

$$(\tilde{\mathbf{X}}\Sigma_X^{1/2})^+ = (\tilde{\mathbf{X}}^+\tilde{\mathbf{X}}\Sigma_X^{1/2})^+(\tilde{\mathbf{X}}\Sigma_X^{1/2}\Sigma_X^{-1/2})^+ = \Sigma_X^{-1/2}\tilde{\mathbf{X}}^+.$$

Using this in (126), we find

$$V_1 + V_2 \lesssim \sigma_\varepsilon^2 \log(n)\cdot p\|\tilde{\mathbf{X}}^+\|^2 = \sigma_\varepsilon^2 \log(n)\frac{p}{\sigma_p^2(\tilde{\mathbf{X}})}.$$

Proof that $\mathbb{P}(\mathcal{B}) \ge 1 - c/n$*:* The bounds $\mathbb{P}(\mathcal{B}_1) \ge 1 - c/n$ and $\mathbb{P}(\mathcal{B}_2) \ge 1 - e^{-cn}$ follow respectively from Theorem 4.6.1 of Vershynin (2019) and Lemma 23 in Appendix C.1 above, by similar reasoning as in the proof of Theorem 16, for example. ■

D.4 Signal to Noise Ratio Bound for Clustered Variables

We present here a lower bound on the signal-to-noise ratio $\xi = \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\|$ in terms of the number $|I_a|$ of features related to cluster a only, for $1 \le a \le K$. We recall the definition

$$I_a \coloneqq \{i \in [p] : \ |A_{ia}| = 1, A_{ib} = 0 \text{ for } b \ne a\}.$$

Lemma 31. $\xi \ge \min_a |I_a| \cdot \lambda_K(\Sigma_Z)/\|\Sigma_E\|$.

Proof For any $v \in \mathbb{R}^K$ with $\|v\| = 1$,

$$\begin{aligned}
v^\top A^\top A v = \|Av\|^2 &= \sum_{i=1}^{p} \left(\sum_{a=1}^{K} A_{ia} v_a \right)^2 \\
&\geq \sum_{i \in I} \left(\sum_{a=1}^{K} A_{ia} v_a \right)^2 \\
&= \sum_{b=1}^{K} \sum_{i \in I_b} A_{ib}^2 v_b^2 \\
&= \sum_{b=1}^{K} |I_b| v_b^2 && (|A_{ib}| = 1 \text{ for } i \in I_b) \\
&\geq \min_a |I_a| \cdot \sum_{b=1}^{K} v_b^2 = \min_a |I_a|. && (\text{since } \|v\| = 1).
\end{aligned}$$

Thus, using $\lambda_K(A\Sigma_Z A^\top) \geq \lambda_K(\Sigma_Z)\lambda_K(A^\top A)$,

$$\xi = \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| \geq \lambda_K(A^\top A)\lambda_K(\Sigma_Z)/\|\Sigma_E\| \geq \min_a |I_a| \lambda_K(\Sigma_Z)/\|\Sigma_E\|,$$

which completes the proof. ■

Appendix E. Properties of the Moore-Penrose Pseudo-Inverse

We state the definition and some properties of the pseudo-inverse in this section for completeness. The material here can be found in Petersen and Pedersen (2012), along with proofs of some of the statements. For a matrix $B \in \mathbb{R}^{n \times m}$, there exists a unique matrix B^+, which we define as the pseudo-inverse of B, satisfying the following four conditions:

$$BB^+B = B \tag{127}$$
$$B^+BB^+ = B^+ \tag{128}$$
$$BB^+ \text{ is symmetric} \tag{129}$$
$$B^+B \text{ is symmetric} \tag{130}$$

We will use the following properties of the pseudo-inverse in this paper.

Lemma 32. *For any $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{m \times d}$,*

$$(BC)^+ = (B^+BC)^+(BCC^+)^+. \tag{131}$$

Furthermore, for any matrix $B \in \mathbb{R}^{n\times m}$ with $r = \text{rank}(B)$ and smallest non-zero singular value $\sigma_r(B)$,

$$B^\top BB^+ = B^\top \tag{132}$$
$$B^\top (BB^\top)^+ = B^+ \tag{133}$$
$$(B^\top B)^+ B^\top = B^+ \tag{134}$$
$$B^+B = I_m \text{ if } r = m \tag{135}$$
$$BB^+ = I_n \text{ if } r = n \tag{136}$$
$$\|B^+\| = 1/\sigma_r(B) \tag{137}$$
$$\text{rank}(B^+) = \text{rank}(B) = r. \tag{138}$$

References

Omar Aguilar and Mike West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18:338–357, 2000.

Theodore W. Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, pages 111–150. University of California Press, 1956.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences USA*, 48(117): 30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 2306–2317, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *In arXiv:1802.01396*, 2018b.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? *In arXiv:1806.09471*, 2018c.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a. doi: 10.1073/pnas.1903070116.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *In arXiv:1903.07571*, 2019b.

Anil Bhattacharya and David B. Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011. doi: 10.1093/biomet/asr013.

Xin Bing, Florentina Bunea, Marten Wegkamp, and Seth Strimas-Mackey. Essential regression. *In arXiv:1905.12696*, 2019.

Xin Bing, Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Prediction in latent factor regression: Adaptive pcr and beyond. *In arXiv:2007.10050*, 2020.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Florentina Bunea, Christophe Giraud, Xi Luo, Martin Royer, and Nicolas Verzelen. Model Assisted Variable Clustering: Minimax-optimal Recovery and Algorithms. *Annals of Statistics*, page to appear, Aug 2019.

Carlos M. Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

Jianqing Fan, Yuan Liao, and Martina Mincheva. High-dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*, 39(6):3320–3356, 12 2011. doi: 10.1214/11-AOS944.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:603–680, 2013a.

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013b.

Jianqing Fan, Lingzhou Xue, and Jiawei Yao. Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292 – 306, 2017.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. *In arXiv:1906.05271*, 2019.

P. Richard Hahn, Carlos M. Carvalho, and Sayan Mukherjee. Partial factor modeling: Predictor-dependent shrinkage for linear regression. *Journal of the American Statistical Association*, 108(503):999–1008, 2013.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *In arXiv:1903.08560*, 2019.

Alan J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Series: Springer Texts in Statistics, 2008.

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

Karl G. Joreskog. Some contributions to maximum likelihood factor analysis. *Psychometrika*, 32:443–482, 1967.

Karl G. Joreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202, 1969.

Karl G. Joreskog. A general method for analysis of covariance structure. *Biometrika*, 57: 239–252, 1970.

Karl G. Joreskog. Factor analysis by least squares and maximum likelihood methods. In A. Ralston K. Enslein and H. S. Wilf, editors, *Statistical Methods for Digital Computers III*, pages 125–153. Wiley, 1977.

Kwang-Sung Jun, Ashok Cutkosky, and Francesco Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. *In arXiv:1905.10680*, 2019.

Derrick N. Lawley. The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, Section A*, 60:64–82, 1940.

Derrick N. Lawley. Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh, Section A*, 61:176–185, 1941.

Derrick N. Lawley. The application of the maximum likelihood method to factor analysis. *British Journal of Psychology*, 33:172–175, 1943.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *In arXiv:1808.00387*, 2018.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. *In arXiv:1712.06559*, 2017.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *In arXiv:1908.05355*, 2019.

Partha P Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation. *In arXiv:1906.03667*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *In arXiv:1903.09139*, 2019.

Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook, 2012.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62:1707–1739, 2009.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460): 1167–1179, 2002a. ISSN 01621459.

James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002b.

James H. Stock and Mark W. Watson. Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2019.

Yue Xing, Qifan Song, and Guang Cheng. Statistical optimality of interpolated nearest neighbor algorithms. *In arXiv:1810.02814*, 2018.