LATENT STRUCTURE IN LINEAR PREDICTION AND CORPORA COMPARISON

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Seth Colin-Bear Strimas-Mackey

August 2022

LATENT STRUCTURE IN LINEAR PREDICTION AND CORPORA COMPARISON

Seth Colin-Bear Strimas-Mackey, Ph.D.

Cornell University 2022

This work first studies the 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 [13] 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.

The second part of this work extends the analysis of the minimum-norm inter-

polating predictor to a larger class of linear predictors of a response $Y \in \mathbb{R}$. Our primary contribution is in establishing finite sample risk bounds for prediction with the ubiquitous Principal Component Regression (PCR) method, under the factor regression model, with the number of principal components adaptively selected from the data—a form of theoretical guarantee that is surprisingly lacking from the PCR literature. To accomplish this, we prove a master theorem that establishes a risk bound for a large class of predictors, including the PCR predictor as a special case. This approach has the benefit of providing a unified framework for the analysis of a wide range of linear prediction methods, under the factor regression setting. In particular, we use our main theorem to recover the risk bounds for the minimum-norm interpolating predictor, and a prediction method tailored to a subclass of factor regression models with identifiable parameters. This model-tailored method can be interpreted as prediction via clusters with latent centers. To address the problem of selecting among a set of candidate predictors, we analyze a simple model selection procedure based on data-splitting, providing an oracle inequality under the factor model to prove that the performance of the selected predictor is close to the optimal candidate.

In the third part of this work, we shift from the latent factor model to developing methodology in the context of topic models, which also rely on latent structure. We provide a new, principled, construction of a distance between two ensembles of independent, but not identically distributed, discrete samples, when each ensemble follows a topic model. Our proposal is a hierarchical Wasserstein distance, that can be used for the comparison of corpora of documents, or any other data sets following topic models. We define the distance by representing a corpus as a discrete measure $\boldsymbol{\theta}$ over a set of clusters corresponding to topics. To a cluster we associate its center, which is itself a discrete measure over

topics. This allows for summarizing both the relative weight of each topic in the corpus (represented by the components of $\boldsymbol{\theta}$) and the topic heterogeneity within the corpus in a single probabilistic representation. The distance between two corpora then follows naturally as a hierarchical Wasserstein distance between the probabilistic representations of the two corpora. We demonstrate that this distance captures differences in the content of the topics between two corpora and their relative coverage. We provide computationally tractable estimates of the distance, as well as accompanying finite sample error bounds relative to their population counterparts. We demonstrate the usage of the distance with an application to the comparison of news sources.

BIOGRAPHICAL SKETCH

Seth Strimas-Mackey grew up in Toronto, Canada, where he studied film and music at Rosedale Heights School of the Arts before shifting to studying engineering physics at the University of Toronto. He first came to Cornell University to start his Ph.D. in physics, which he pursued for two years. He then switched to begin his Ph.D. in statistics in the Department of Statistics and Data Science at Cornell after developing an interest in the field. He is very fortunate to be co-advised by Professors Florentina Bunea and Marten Wegkamp.

I dedicate this thesis to my brother Matthew.

ACKNOWLEDGEMENTS

In my time at Cornell I've enjoyed the support and companionship of some incredible people. Nearing the end of my time here, I'd like to take the chance to express my gratitude.

I first met my advisors Florentina Bunea and Marten Wegkamp when seeking to transfer from the Ph.D. program in physics at Cornell to my Ph.D. in statistics. Although I had limited background in statistics, they took a chance to accept me as their student, an opportunity for which I will be forever grateful. Throughout my Ph.D. they taught me about every aspect of the research process. I am extremely thankful to them for their endless patience, kindness, and the enormous support they have provided me.

I am also indebted to my parents and siblings for their constant support. Due to a happy coincidence, my brother Matt moved to Ithaca just a couple years after me to work at Cornell. Having him live nearby has been one of the great joys of my time here.

Lastly, I'd like to thank Geethika, Edward, Peter, Daniel, Ben, Fiona, and all the other friends I've met at Cornell. They've made my time here a great experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTERPOLATING PREDICTORS IN HIGH-DIMENSIONAL FACTOR REGRESSION

1.1 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 [13–18,44,50,61,72,73,75,76,79,93].

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 $\widehat{\alpha} = \boldsymbol{X}^+\mathbf{y}$ is based on the Moore-Penrose pseudo-inverse of the $n \times p$ data matrix $\boldsymbol{X}$ 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 $\boldsymbol{X}$ has full rank, interpolates the data. The interpolation property of $\widehat{\alpha}$ means that $\boldsymbol{X}\widehat{\alpha} = \mathbf{y}$. 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 1.2, 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 [50].

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 [13] 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 [13], 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 (1.7) of Section 1.3.1 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 1.3 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 1.4 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 13 in Section 1.4.2. It provides a finite-sample bound on the *excess risk* $R(\widehat{\alpha}) - \sigma^2$ of $\widehat{\alpha}$ in the high-dimensional setting $p > n$, relative to the natural risk benchmark $\mathbb{E}[\varepsilon^2] := \sigma^2$ 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 [50], but does indeed decrease to zero, as desired. We remark that at least for Gaussian (X, y), [13] provides an alternative bound to Theorem 13. However, Theorem 13 provides an improved rate for typical factor regression models, and in particular provides examples when the upper bound on the excess risk in [13] diverges, yet our results show that prediction is consistent; see Section 1.4.3 for a detailed comparison.

Table 1.1 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

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

Table 1.1: 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).

if $p \gg n \log n$ 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 1.3.1, 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 1.3.1, 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^2. As argued in Section 1.3.4, 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 1.3.3 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 [57–60, 69–71], including the pertinent work [4]. A number of works ranging from purely Bayesian [1, 19, 35, 49] to variational Bayes [30] to frequentist [25, 40–43, 55, 56, 86–88] 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 [25, 40–43], 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 1.4.4 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 1.2 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 1.3 introduces the factor regression model (1.5) 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 (1.5) are given in Section 1.3.1, 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 1.3.2.

Section 1.3.3 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 (1.5), $\|\alpha^*\| \to 0$ 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 1.3.4, commenting on assumptions on the operator norm of Σ_X, and on implications for prediction with the GLS.

The remainder of the paper, Section 1.4, contains our analysis of the GLS $\widehat{\alpha}$ and its prediction risk, under the factor regression model. Section 1.4.1 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 1.4.2 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 1.4.1, $R(\widehat{\alpha})/R(\mathbf{0})$ does not approach 1. Instead, the finite-sample risk bound in Theorem 13 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 1.4.3 presents a comparison with recent related work. In particular, we give a detailed comparison with [13], 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 [13] under the factor regression model, which are derived in Appendix A.3.4. Table 1.2 of Section 1.4.3 summarizes our findings that the bound on the excess risk in [13] is often larger in order of magnitude than the bound given in Theorem 13 of Section 1.4.2. In particular, we exhibit instances of the factor regression model class under which the excess risk upper bound in [13] diverges, yet our upper bound approaches zero. We also compare our work to [75], 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 1.4.4 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 37 in the Appendix complements Theorem 13 by showing the risk behavior of $\widehat{\alpha}$ for $n > c \cdot p$ for an absolute constant $c > 0$, and is included for completeness.

1.1.1 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 A.5 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.

1.2 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 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the corresponding data matrix with rows $X_1, \ldots X_n$, and let $\mathbf{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. 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

$$\widehat{\alpha} := \arg\min \left\{ \|\alpha\| : \ \|\boldsymbol{X}\alpha - \mathbf{y}\| = \min_{u} \|\boldsymbol{X}u - \mathbf{y}\| \right\}. \tag{1.1}$$

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

$$R(\alpha) := \mathbb{E}_{X,y}[(X^\top \alpha - y)^2]. \tag{1.2}$$

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

$$\widehat{\alpha} := \arg \min_{\alpha:\ \boldsymbol{X}\alpha = \mathbf{y}} \|\alpha\|. \tag{1.3}$$

Regardless of the rank of $\boldsymbol{X}$, Equation (1.1) always has the closed form solution $\widehat{\alpha} = \boldsymbol{X}^+ \mathbf{y}$, where $\boldsymbol{X}^+$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$; we prove this fact in section A.4.1 for completeness. We begin our consideration of the minimum-norm estimator $\widehat{\alpha} = \boldsymbol{X}^+ \mathbf{y}$ 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.1. 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 $1 - ce^{-c'n}$ for absolute constants*

$c, c' > 0$,

$$\left| \frac{R(\widehat{\alpha})}{R(\mathbf{0})} - 1 \right| \lesssim \sqrt{\frac{n}{\mathrm{r_e}(\Sigma_X)}}. \tag{1.4}$$

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 [50] depicts an example of this behavior: it plots $\mathbb{E}[\|\widehat{\alpha} - \alpha\|^2 | \boldsymbol{X}]$ 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 [13] 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 (1.5) given below. In this model, while $\mathrm{r_e}(\Sigma_X)/n$ remains bounded, $\|\Sigma_X\|$ typically grows with p (see Lemma 3 below), in contrast to the assumption $\|\Sigma_X\| = 1$ of the definition of benign matrices in [13]. Furthermore, the results in [13] only apply to model (1.5) 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 1.4.3 below. We find that in this common regime, we obtain a tighter bound.

1.3 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

$$X = AZ + E, \qquad y = Z^\top \beta + \varepsilon, \tag{1.5}$$

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 $\Sigma_E := \mathrm{Cov}(E)$, $\Sigma_Z := \mathrm{Cov}(Z)$ and $\sigma^2 := \mathrm{Var}(\varepsilon)$. For the remainder of the paper we will assume that the data consist of n i.i.d. pairs (X_i, y_i) satisfying (1.5), in that

$$X_i = AZ_i + E_i, \qquad y_i = Z_i^\top \beta + \varepsilon_i \qquad \forall i \in [n], \tag{1.6}$$

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 $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ 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 $\boldsymbol{Z} \in \mathbb{R}^{n \times K}$ 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 1.3.1. The risk benchmarks are introduced and analyzed in Section 1.3.2. Section 1.3.3 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 1.3.4.

1.3.1 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 (1.5), we find

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

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

$$\frac{\mathrm{r_e}(\Sigma_X)}{n} \le \frac{K}{n} + \frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}, \tag{1.7}$$

where

$$\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\|, \tag{1.8}$$

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 [4], $\Sigma_E = I_p$, 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 (1.7).

Lemma 2. *Under model (1.5), we have* $\mathrm{r_e}(\Sigma_X)/n \le c_3$ *whenever*

$$\frac{K}{n} \le c_1 \quad \text{and} \quad \xi \ge c_2 \frac{\mathrm{r_e}(\Sigma_E)}{n}, \tag{1.9}$$

for positive absolute constants c_1, c_2, c_3.

Remark 1. We remark on conditions under which (1.9) holds. 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,

$$\mathrm{r_e}(\Sigma_E) \asymp p, \quad \text{and} \quad \xi = \frac{\lambda_K(A\Sigma_Z A^\top)}{\|\Sigma_E\|} \asymp \lambda_K(A^\top A), \tag{1.10}$$

so the condition (1.9) reduces to $K/n \le c_1$ and

$$\lambda_K(A^\top A) \gtrsim \frac{p}{n}. \tag{1.11}$$

We give a few examples of A that imply (1.11):

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 (1.11) 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 [91], for example) we once again have $\lambda_K(A^\top A) \gtrsim p$, with high probability, and (1.11) holds.

3. In other situations, (1.11) is an assumption. It is a very natural, and mild, requirement in factor regression models, and if A is structured and sparse, (1.11) can be given further interpretation. For instance, the model $X = AZ+E$ has been used and analyzed in [32] 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 38 in Appendix A.4.4) that $\lambda_K(A^\top A) \ge m$. Furthermore, when these K clusters are approximately balanced, then $m \approx p/K$ and (1.11) holds, provided $K \lesssim n$.

The positive repercussion of Lemma 2 is that under condition (1.9) 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 (1.5) with restrictions (1.9), 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 1.3.2, and show that $R(\widehat{\alpha})$ can indeed approach these benchmarks in Sections 1.4.1 and 1.4.2.

For completeness, we offer the following result characterizing the spectrum of Σ_X under the factor regression model. In particular, as announced in Section 1.2, we find that the operator norm $\|\Sigma_X\|$ diverges with p under mild conditions. The proof can be found in Appendix A.2.1.

Lemma 3. *Suppose that for some $c_1, c_2, C_1, C_2 \in (0,\infty)$,*

$$c_1 \le \lambda_K(\Sigma_Z) \le \|\Sigma_Z\| \le C_1 \quad \text{and} \quad c_2 < \lambda_p(\Sigma_E) \le \|\Sigma_E\| < C_2. \tag{1.12}$$

The spectrum of Σ_X can then be characterized as follows:

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

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

3. $c_2 \leq \lambda_i(\Sigma_X) \leq C_2$ *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 1.3.3, and use this in Section 1.3.4 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.

1.3.2 Risk Benchmarks

We introduce here two natural benchmarks for $R(\widehat{\alpha})$ under the factor regression model, and characterize their relationship. Under model (1.5), if $Z \in \mathbb{R}^K$ were observed, the optimal risk of a linear oracle with access to Z is

$$\min_{v \in \mathbb{R}^K} \mathbb{E}\left[(Z^\top v - y)^2\right] = \mathbb{E}[\varepsilon^2] = \sigma^2, \tag{1.13}$$

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

$$\alpha^* \in \arg\min_{\alpha \in \mathbb{R}^p} R(\alpha). \tag{1.14}$$

Lemma 34 in Appendix A.4 shows that for arbitrary zero-mean (X, y) with finite second moments, $\alpha^* = \Sigma_X^+ \Sigma_{XY}$ is a minimizer of $R(\alpha)$, where $\Sigma_{Xy} := \mathbb{E}[Xy] \in \mathbb{R}^p$ is

the vector of component-wise covariances.

We can characterize the difference between these two benchmarks, σ^2 and $R(\alpha^*)$, as follows. See Appendix A.2.2 for the proof of this result.

Lemma 4 (Comparison of risk benchmarks)**.** *Suppose model (1.5) holds and let ξ be the signal-to-noise ratio defined in (1.8). We have*

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

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

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 $R(\alpha^*) - \sigma^2 \to 0$, as $p \to \infty$.

Although the optimal risk $R(\alpha^*)$ is always greater than the oracle risk σ^2 (part 1 of Lemma 4), the bound $\|\beta\|_{\Sigma_Z}^2/\xi$ on the difference $R(\alpha^*) - \sigma^2$ in part 2 of Lemma 4 is not a leading term in the excess risk bound given in Theorem 13. 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 7 in the next section. This is an important feature of the FRM, and its repercussions are discussed in Section 1.3.4.

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

In this section we investigate the properties of the population-level predictor α^*, defined in (1.14), under the factor regression model (1.5). In particular, we prove that $\|\alpha^*\| \to 0$ and yet $R(\mathbf{0}) - R(\alpha^*) > 0$ under the conditions

$$\lim_{p\to\infty} \|\beta\|_{\Sigma_Z}^2 / \lambda_K(A\Sigma_Z A^\top) = 0 \text{ and } \liminf_{p\to\infty} \|\beta\|_{\Sigma_Z} > 0. \tag{1.15}$$

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 A.2.3. We first characterize the norms $\|\alpha^*\|$ and $\|\alpha^*\|_{\Sigma_X}$; the latter norm is of interest via the identity

$$R(\mathbf{0}) - R(\alpha^*) = \|\alpha^*\|_{\Sigma_X}^2. \tag{1.16}$$

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

$$\alpha^{*\top}X = (A^\top A^{+\top}\beta)^\top Z = \beta^\top Z, \tag{1.17}$$

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 5. *Suppose model (1.5) holds, that $\Sigma_E = 0$, and that Σ_Z and A are full rank. Then, $\alpha^* = A^{+\top}\beta$, and*

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

We next find that in the more realistic case, when $\Sigma_E \neq 0$, even though identity (1.17) 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 6. *Suppose model (1.5) 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 $\kappa(\Sigma_E) < C < \infty$,*

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

Remark 2. We illustrate our findings in Lemmas 5 and 6 with the following example (that we will use in our simulations in Section 1.4.4), 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,

$$\alpha^* = \frac{\sigma_Z^2}{\sigma_E^2 + a^2\sigma_Z^2} A\beta \tag{1.18}$$

$$\|\alpha^*\|^2 = \frac{a^2\sigma_Z^2}{(\sigma_E^2 + a^2\sigma_Z^2)^2} \|\beta\|^2_{\Sigma_Z} \tag{1.19}$$

$$\|\alpha^*\|^2_{\Sigma_X} = \frac{a^2\sigma_Z^2}{\sigma_E^2 + a^2\sigma_Z^2} \|\beta\|^2_{\Sigma_Z}. \tag{1.20}$$

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 $\|\beta\|^2_{\Sigma_Z}/\lambda_K(A\Sigma_Z A^\top) \to 0$ forces $\|\alpha^*\| \to 0$, while at the same time $\|\alpha^*\|^2_{\Sigma_X} \asymp \|\beta\|^2_{\Sigma_Z}$ when ξ is bounded below (in fact, $\|\alpha^*\|^2_{\Sigma_X}/\|\beta\|^2_{\Sigma_Z} \to 1$ 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 (1.18) 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 1.4.4.

Identity (1.16), Lemma 5 and Lemma 6 imply the following conclusion.

Corollary 7. *Suppose model (1.5) 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 (1.5), $\Sigma_E = 0$ and A, Σ_Z are full rank. Then, in either case, condition (1.15) 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.

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

The model (1.5) 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 (1.5) to a model in which X continues to follow a factor model, but y is connected to X via a linear model:

$$X = AZ + E, \qquad y = X^\top \theta + \eta, \tag{1.21}$$

where $\theta \in \mathbb{R}^p$ is a generic p-dimensional regression vector, and η is zero-mean noise independent of X. Model (1.21) 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 36 in Appendix A.4.2 shows the simple fact that if the factor regression model (1.5) holds, then (1.21) 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 (1.5)*, which we studied the properties of in Section 1.3.3 above.

We can thus compare model (1.5) and (1.21) 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 1.3.3 we found that $\|\alpha^*\| \to 0$, provided (1.15) holds. Thus, when the factor regression model (1.5) is viewed as a particular case of (1.21), 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 (1.21), 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 [13] 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 [13] contains a bias term given by

$$\|\theta\|^2 \|\Sigma_X\| \max\left\{\sqrt{\frac{\mathrm{r_e}(\Sigma_X)}{n}}, \frac{\mathrm{r_e}(\Sigma_X)}{n}\right\}. \tag{1.22}$$

We examine this bound assuming further that model (1.21) holds. Since

$$\|\Sigma_X\| \max\left\{\sqrt{\frac{\mathrm{r_e}(\Sigma_X)}{n}}, \frac{\mathrm{r_e}(\Sigma_X)}{n}\right\} = \max\left\{\sqrt{\frac{\|\Sigma_X\|\mathrm{tr}(\Sigma_X)}{n}}, \frac{\mathrm{tr}(\Sigma_X)}{n}\right\} \geq \frac{\mathrm{tr}(\Sigma_X)}{n} \tag{1.23}$$

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 (1.21) with mild assumptions on either Σ_E (e.g., $\Sigma_E \asymp I_p$) or A (see Remark 1), the bias term (1.22) will only converge to zero if $\|\theta\| \to 0$.

As noted above, $\|\theta\| \to 0$ is rather unnatural in a generic model (1.21). However, we also noted that when (X, y) are Gaussian and the factor regression model (1.5) holds, then (1.21) holds with $\|\theta\| = \|\alpha^*\| \to 0$, which means that the bias term (1.22) can converge to zero when the data is generated by model (1.5). We take this as indication that the bias in prediction with $\widehat{\alpha}$ can be significantly lower in the factor regression model (1.5) compared to a generic model (1.21) as a result of the joint low-dimensional structure of model (1.5).

We note that this discussion is only based on an upper bound (1.22) on the bias term of the prediction risk. It nevertheless motivates a full investigation of an alternative upper bound to (1.22), directly derived under model (1.5). This is the subject of Section 1.4 below, with our main result presented in Theorem 13.

Remark 3. The authors of [13] take a different route, complementary to ours, in their analysis of the bound (1.22). 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 [13] 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 (1.22) can converge to zero even when $||\theta|| \not\to 0$, for a generic θ.

However, as we have shown in Lemma 3 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 (1.22) will vanish, in the regime $\mathrm{r_e}(\Sigma_X)/n \to 0$, even though $||\Sigma_X|| \to \infty$. The results of Section 1.4 can thus be viewed as complementary to those in [13].

1.4 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 1.3.2.

1.4.1 Exact Adaptation in Factor Regression Models with Noiseless Features

We begin our analysis by considering an extreme case of model (1.5), 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 A.3.1. 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 $\varepsilon = \sigma\tilde{\varepsilon}$, 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 5 above, we showed that $\|\alpha^*\|^2 = \beta^\top(A^\top A)^{-1}\beta$ when $\Sigma_E = 0$, and as a result, Corollary 7 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 $(\sigma^2 \log n)/\lambda_K(A\Sigma_Z A^\top) \to 0$ as $n \to \infty$.

Lemma 8. *Under model (1.5) 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 $c > 0$,*

$$\|\widehat{\alpha}\|^2 \lesssim \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\left(\|\beta\|^2_{\Sigma_Z} + \sigma^2\frac{K\log n}{n}\right). \tag{1.24}$$

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 7. 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 $\widehat{y}_z := Z^\top \widehat{\beta}$ be the predictor based on the least-squares regression coefficients $\widehat{\beta} := \boldsymbol{Z}^+ \mathbf{y}$ of $\mathbf{y}$ onto $\boldsymbol{Z}$; this is the classical least-squares prediction of y under model (1.5) that an oracle would use if it had access to the unobserved data matrix $\boldsymbol{Z}$, 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 $(\boldsymbol{X}, \mathbf{y})$ only. Theorem 9.1 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 9 (Factor regression with noiseless features)**.** *Under model (1.5) with* $\Sigma_E = 0$*, suppose that Assumption 2 holds.*

1. *Then, on the event that the matrix* $\boldsymbol{Z}$ *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* $c > 0$*,* $\boldsymbol{Z}$ *has full rank* K *and*

$$R(\widehat{\alpha}) - \sigma^2 \lesssim \sigma^2 \frac{K \log n}{n} \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\varepsilon}}[R(\widehat{\alpha})] - \sigma^2 \gtrsim \sigma^2 \frac{K}{n}. \tag{1.25}$$

The risk bounds (1.25) 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 (1.5). We note that, since $\operatorname{rank}(\boldsymbol{X}) = K < n$, $\mathbf{y}$ may not lie in the range of $\boldsymbol{X}$ and so $\widehat{\alpha}$ may not interpolate. Nonetheless, under model (1.5), 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., [86], 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 $\boldsymbol{X}$ to estimate a vector of coefficients $(\boldsymbol{X}\widehat{U}_K)^+\mathbf{y}$. Here $\widehat{U}_K \in \mathbb{R}^{p\times K}$ has columns equal to the first K eigenvectors of the sample covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}/n$. A new response y is then predicted by $\widehat{\alpha}_{\mathrm{PCR}}^\top X$, where $\widehat{\alpha}_{\mathrm{PCR}} := \widehat{U}_K(\boldsymbol{X}\widehat{U}_K)^+\mathbf{y}$ and X is the new feature vector. The following lemma states that the PCR and GLS predictors coincide when $\Sigma_E = 0$.

Lemma 10. *Define* $\widehat{\alpha}_{\mathrm{PCR}} := \widehat{U}_K(\boldsymbol{X}\widehat{U}_K)^+\mathbf{y}$. *On the event* $\{\mathrm{rank}(\boldsymbol{X}) = K\}$, $\widehat{\alpha} = \widehat{\alpha}_{\mathrm{PCR}}$. *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 $\Sigma_E = 0$.

1.4.2 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 1.3.2 above, under the factor regression model (1.5) with $E \neq 0$.

Our main result, Theorem 13 below, shows that despite the fact that $\widehat{\alpha}$ interpolates, in that $\boldsymbol{X}\widehat{\alpha} = \mathbf{y}$ (Proposition 11), and that $\|\widehat{\alpha}\| \to 0$ (Lemma 12), the excess risks can vanish as a result of approximate adaptation to the embedded low-dimensional structure of (1.5). The estimator $\widehat{\alpha}$ is guaranteed to interpolate the data whenever $\text{rank}(\boldsymbol{X}) = n$, or equivalently, the smallest singular value $\sigma_n(\boldsymbol{X}) > 0$. 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 A.3.2.

Proposition 11. *Under model (1.5), 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* $c > 0$*,*

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

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

General existing bounds of the type $\sigma_n(\mathbf{X}) \gtrsim (\sqrt{p} - \sqrt{n})$ are by now well established in random matrix theory [85]. When $p > C \cdot n$ for some $C > 1$ and the entries of $\boldsymbol{X}$ are i.i.d. sub-Gaussian with zero mean and unit variance, Theorem 1.1 in [85] implies that $\sigma_n^2(\mathbf{X}) \gtrsim p$ with high probability. By comparison, Proposition 11 holds for $\boldsymbol{X}$ 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 8, $\|\widehat{\alpha}\| \to 0$, mimicking the behavior of the best linear predictor α^*. We proved in Lemma 6 and Corollary 7 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 11, which requires that the effective rank $\mathrm{r_e}(\Sigma_E)$ is larger than a constant times n.

Lemma 12. *Under model (1.5), 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 $c > 0$,*

$$\|\widehat{\alpha}\|^2 \lesssim \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\|\beta\|^2_{\Sigma_Z} + \sigma^2 \frac{n \log n}{\mathrm{r_e}(\Sigma_E)}. \tag{1.26}$$

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 9 above showed that $R(\widehat{\alpha}) - \sigma^2$ 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 13 (Main result: Risk bound for factor regression)**.** *Under model (1.5), 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 $c > 0$,*

$$R(\widehat{\alpha}) - R(\alpha^*) \leq R(\widehat{\alpha}) - \sigma^2 \lesssim \frac{\|\beta\|^2_{\Sigma_Z}}{\xi} \cdot \frac{\mathrm{r_e}(\Sigma_E)}{n} + \sigma^2 \frac{n \log n}{\mathrm{r_e}(\Sigma_E)} + \sigma^2 \frac{K \log n}{n}. \tag{1.27}$$

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

Remark 4. Suppose $n \gg \sigma^2 K \log n$ and $\mathrm{r_e}(\Sigma_E) \gg \sigma^2 n \log n$. We then find that $\widehat{\alpha}$ interpolates by Proposition 11, 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 12 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 $R(\widehat{\alpha}) - \sigma^2 \to 0$. Indeed, this assumption, together with $n \gg \sigma^2 K \log n$ and $\mathrm{r_e}(\Sigma_E) \gg \sigma^2 n \log n$, ensures that the right-hand side of the inequality (1.27) in Theorem 13 is asymptotically negligible.

The first inequality in (1.27) is an immediate consequence of the first part of Lemma 4 above. We now discuss the three terms appearing in the upper bound (1.27) of Theorem 13. A comparison with the risk bound in Theorem 9 above, where the feature noise E is equal to zero, reveals that the term $\sigma^2 K \log(n)/n$ in (1.27) 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,

$$\mathrm{r_e}(\Sigma_E) \asymp p, \quad \text{and} \quad \xi = \frac{\lambda_K(A\Sigma_Z A^\top)}{\|\Sigma_E\|} \gtrsim \lambda_K(A^\top A). \tag{1.28}$$

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 (1.27) can be bounded as

$$B_Z := \frac{\|\beta\|^2_{\Sigma_Z}}{\xi} \cdot \frac{\mathrm{r_e}(\Sigma_E)}{n} \lesssim \frac{Kp}{n \cdot \lambda_K(A^\top A)}; \tag{1.29}$$

it thus approaches zero whenever

$$\lambda_K(A^\top A) \gg \frac{Kp}{n}. \tag{1.30}$$

We mention that the examples of A in Remark 1 of Section 1.3.1 all imply (1.30), 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 14 below.

Corollary 14. *Under the same conditions as in Theorem 13, 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 $c > 0$,*

$$R(\widehat{\alpha}) - R(\alpha^*) \le R(\widehat{\alpha}) - \sigma^2 \lesssim \frac{K}{\lambda_K(A^\top A)} \times \frac{p}{n} + \sigma^2 \left(\frac{n}{p} + \frac{K}{n}\right) \log n. \tag{1.31}$$

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

$$R(\widehat{\alpha}) - R(\alpha^*) \le R(\widehat{\alpha}) - \sigma^2 \lesssim \frac{K^2}{n} + \sigma^2 \left(\frac{n}{p} + \frac{K}{n}\right) \log n. \tag{1.32}$$

Figure 1.1 illustrates the risk behavior proved in Theorem 13. 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 (1.5), in Appendix A.4.3.

1.4.3 Comparison to Existing Results

The recent paper [13] 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, [13] does not consider model (1.5), and in fact assumes $\mathbb{E}[y|X] = X^\top \theta$ for some $\theta \in \mathbb{R}^p$, which is typically not satisfied under (1.5) when (X, y) are sub-Gaussian, but not Gaussian.

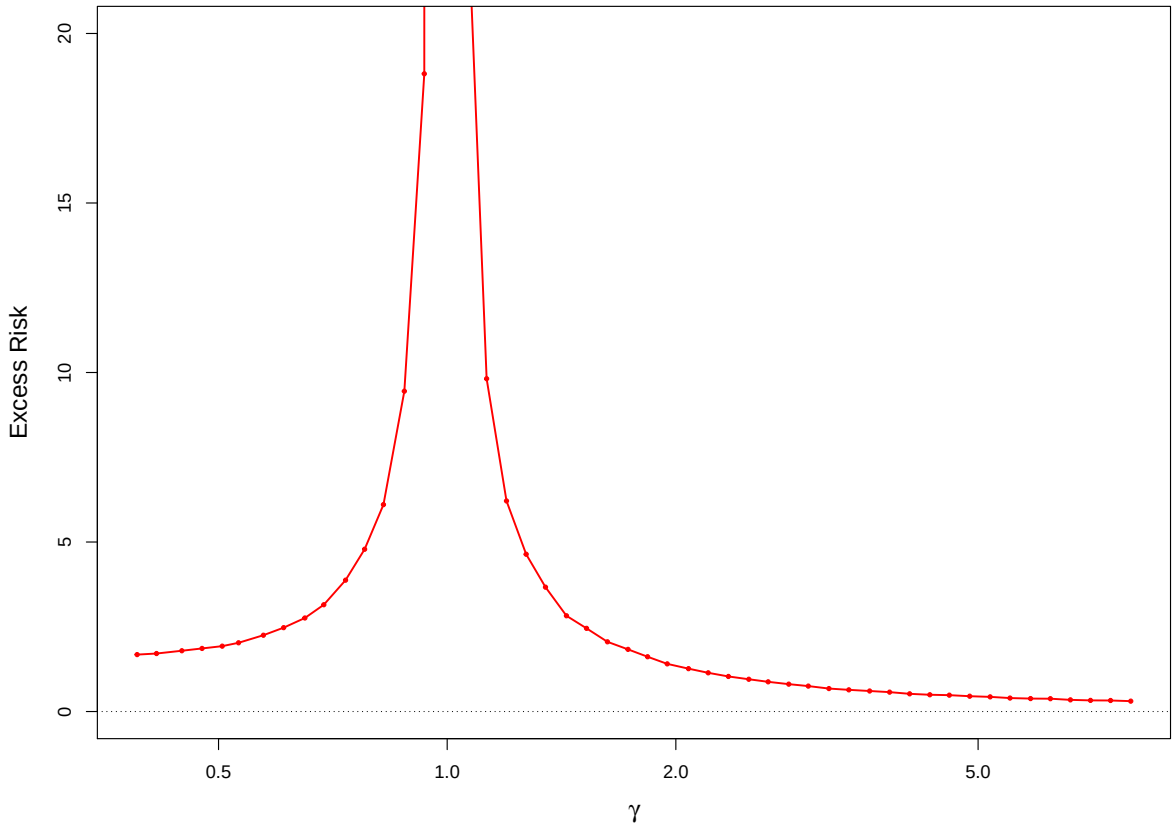

Figure 1.1: Excess prediction risk $R(\widehat{\alpha}) - \sigma^2$ of the minimum-norm predictor under the factor regression model as a function of $\gamma = p/n$. Here K increases linearly from 16 to 64, $n = \lfloor K^{1.5} \rfloor$ and thus increases from 64 to 512, and p increases from 33 to 4066. 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.

When the data are jointly Gaussian this assumption is, however, satisfied under model (1.5). For this common case, Table 1.2 compares the respective bounds on the bias and variance terms corresponding to our Theorem 13 and Theorem 4 of [13], respectively. Again, we emphasize that the results from [13] do not hold in general for our modeling setup, but can be used to obtain the bounds in Table 1.2 in the Gaussian case. The entries in the second column of Table 1.2 correspond to the bias in [13] under model (1.5), simplified in this table

Regime	Bias in Theorem 13	Bias in Theorem 4 of [13]	Common variance
$p \geq n \cdot \xi$	$\|\beta\|^2_{\Sigma_Z} \cdot p/(n \cdot \xi)$	$\|\beta\|^2_{\Sigma_Z} \cdot p/(n \cdot \xi)$	$\sigma^2 \log n \{(n/p) + (K/n)\}$
$p \ll n \cdot \xi$	$\|\beta\|^2_{\Sigma_Z} \cdot p/(n \cdot \xi)$	$\|\beta\|^2_{\Sigma_Z} \cdot \sqrt{p/(n \cdot \xi)}$	
$\xi \approx p, \|\beta\|^2_{\Sigma_Z} \approx K$	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 1.2: Comparison of risk bounds for Gaussian data.

for ease of comparison.[1]

In the setting of this comparison, the variance terms in our Theorem 13 and the bound in [13] have the same rate, which we display in the third column of Table 1.2. From the first row of Table 1.2 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 1.2 shows that the bias in [13] becomes $\|\beta\|^2_{\Sigma_Z} \sqrt{p/(n \cdot \xi)}$, which is larger than our bias bound in Theorem 13 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 [13] can diverge while our bound in Theorem 13 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 [13] is not always tight.

The third row of Table 1.2 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 [13]

[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 $c \in (0, 1)$.

diverges. Further details and discussion on the comparison of these two results are deferred to Appendix A.3.4.

A latent factor regression model similar to (1.5) has also been studied in Section 7 of [75] for the ridge regression estimator that minimizes the fit $\|\mathbf{y} - \boldsymbol{X}a\|^2 + \lambda\|a\|^2$ for any $\lambda > 0$ (strict). Their model is a particular case of our model (1.5), 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 [75] 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 [75] 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.

1.4.4 Comparison to Other Predictors

In Lemma 10 of Section 1.4.1 above we showed that in the case of noiseless features, when $\Sigma_E = 0$, the regression vector $\widehat{\alpha}_{\text{PCR}}$ obtained by PCR is exactly equal to the GLS regression vector $\widehat{\alpha}$ on the event $\{\text{rank}(\boldsymbol{Z}) = K\}$, 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 $\widetilde{\alpha}_{\text{PCR}} := U_K(\boldsymbol{X}U_K)^+\mathbf{y}$ of PCR under the factor regression model setting (1.5) 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 A.3.3, of the following risk bound for $R(\widetilde{\alpha}_{\mathrm{PCR}})$. For a detailed comparison of PCR and the GLS, see [20], which analyzes the PCR predictor with the empirical matrix $\widehat{U}_{\widehat{K}}$, for a new, data adaptive, estimator $\widehat{K}$ of K.

Theorem 15. *Under model (1.5), 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 $1 - c/n$,*

$$R(\widetilde{\alpha}_{\mathrm{PCR}}) - \sigma^2 \lesssim \|\Sigma_E\| \cdot \|\alpha^*\|^2 \frac{p}{n} + R(\alpha^*) \frac{K \log(n)}{n} \tag{1.33}$$

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

$$R(\widetilde{\alpha}_{\mathrm{PCR}}) - \sigma^2 \lesssim \sigma_\varepsilon^2 \frac{K \log(n)}{n} \tag{1.34}$$

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

$$R(\widetilde{\alpha}_{\mathrm{PCR}}) - \sigma^2 \lesssim \kappa(\Sigma_E) \frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \frac{p}{n} + \sigma^2 \frac{K \log n}{n}, \tag{1.35}$$

where $\kappa(\Sigma_E) := \lambda_1(\Sigma_E)/\lambda_p(\Sigma_E)$ 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 13 for the GLS $\widehat{\alpha}$ takes the form

$$R(\widehat{\alpha}) - \sigma^2 \lesssim \frac{\|\beta\|_{\Sigma_Z}^2}{\xi} \frac{p}{n} + \sigma^2 \log n \left(\frac{K}{n} + \frac{n}{p} \right). \tag{1.36}$$

The additional term $\sigma^2 n \log n / p$ in this bound is absent in the PCR prediction bound (1.35) 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

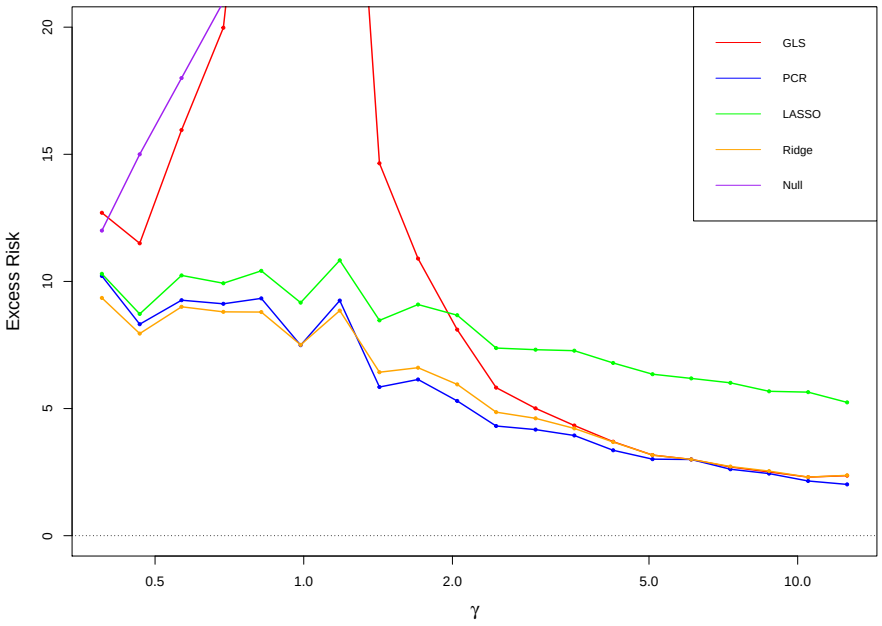

Figure 1.2: Excess prediction risk of GLS, PCR, LASSO, Ridge regression, and the null predictor as a function of $\gamma = p/n$. Here K increases linearly from 12 to 69, $n = \lfloor K^{1.5} \rfloor$ and thus increases from 41 to 573, and p increases from 16 to 7215. Further, $\Sigma_E = I_p$, $\Sigma_Z = I_K$, $\beta = (1, \ldots, 1)^\top$, and A is generated by sampling each entry iid from $N(0, 1/\sqrt{K})$.

problem setting. This is despite the fact that the GLS interpolates the data (when $\text{rank}(\boldsymbol{X}) = n$) 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 1.2 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

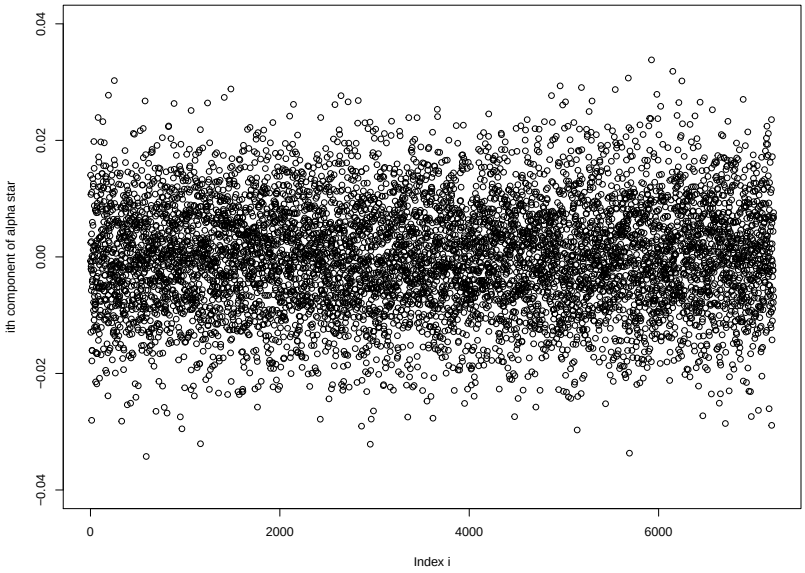

Figure 1.3: A scatter plot of the components of α^*, from the point in the simulation of Figure 1.2 with the largest value of γ. Here $p = 7215$, $K = 69$, $\Sigma_E = I_p$, $\Sigma_Z = I_K$, and A is generated by sampling each entry iid from $N(0, 1/\sqrt{K})$.

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 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 [50, 75].

We plot the coefficients of α^* in Figure 1.3 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 (1.18) of Remark 2,

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

Figure 1.4 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 not necessarily be sparse or even approximately sparse, and we should expect the GLS to outperform the LASSO (see Remark 2 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 1.2 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.

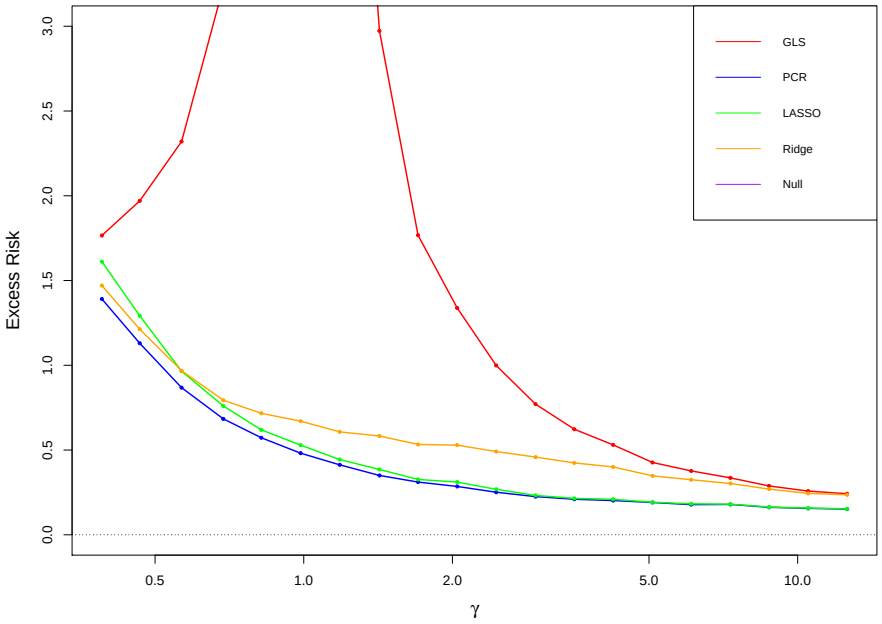

Figure 1.4: Excess prediction risk of GLS, PCR, LASSO, Ridge regression, and the null predictor as a function of $\gamma = p/n$. Null risk is not visible on plot since it is larger than the maximum plotted value. Here K increases linearly from 12 to 69, $n = \lfloor K^{1.5} \rfloor$ and thus increases 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 $\sqrt{p}$.

CHAPTER 2

PREDICTION UNDER LATENT FACTOR REGRESSION: ADAPTIVE PCR, INTERPOLATING PREDICTORS AND BEYOND

2.1 Introduction

This work is devoted to the derivation and analysis of finite sample prediction risk bounds for a class of linear predictors of a random response $Y \in \mathbb{R}$ from a high-dimensional, and possibly highly correlated random vector $X \in \mathbb{R}^p$, when the vector (X, Y) follows a latent factor regression model, generated by a latent vector of dimension lower than p. We assume that there exist a random, unobservable, latent vector $Z \in \mathbb{R}^K$, a deterministic matrix $A \in \mathbb{R}^{p \times K}$, and a coefficient vector $\beta \in \mathbb{R}^K$ such that

$$\begin{aligned} Y &= Z^\top \beta + \varepsilon, \\ X &= AZ + W, \end{aligned} \tag{2.1}$$

with some unknown $K < p$. The random noise $\varepsilon \in \mathbb{R}$ and $W \in \mathbb{R}^p$ have mean zero and second moments $\sigma^2 := \mathbb{E}[\varepsilon^2]$ and $\Sigma_W := \mathbb{E}[WW^\top]$, respectively. The random variable ε and random vectors W and Z are mutually independent. Throughout the paper, both $\Sigma_Z := \mathbb{E}[ZZ^\top]$ and A have rank equal to K.

Independently of this model formulation, but based on the belief that Y depends chiefly on a lower-dimensional approximation of X, prediction of Y via principal components (PCR) is perhaps the most utilized scheme, with a history dating back many decades [53, 64]. Given the data $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ consisting of n independent copies of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, PCR-$k$

predicts $Y_* \in \mathbb{R}$ after observing a new data point $X_* \in \mathbb{R}^p$ by

$$\begin{aligned} \widehat{Y}^*_{\mathbf{U}_k} &= X_*^\top \mathbf{U}_k \left[\mathbf{U}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \mathbf{U}_k\right]^+ \mathbf{U}_k^\top \boldsymbol{X}^\top \boldsymbol{Y} \\ &= X_*^\top \mathbf{U}_k \left[\boldsymbol{X}\mathbf{U}_k\right]^+ \boldsymbol{Y}, \end{aligned} \tag{2.2}$$

where $\mathbf{U}_k$ is the $p \times k$ matrix of the top eigenvectors of the sample covariance matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$, relative to the largest k eigenvalues, where k is ideally determined in a data-dependent fashion and M^+ denotes the Moore-Penrose inverse of a matrix M.

Model (2.1) provides a natural context for the theoretical analysis of PCR-k prediction. It is perhaps surprising that its theoretical study so far is limited to asymptotic analyses of the out-of-sample prediction risk for PCR-K as $p, n \to \infty$ [10, 86], and finite sample / asymptotic risk bounds on the in-sample prediction accuracy of PCR-K [8, 12, 42, 43, 63] in identifiable factor models with known and fixed K.

To the best of our knowledge, finite sample prediction risk bounds for $\widehat{Y}^*_{\mathbf{U}_k}$, corresponding to data-dependent choices of k, are lacking in the literature, and their study under factor models of unknown K, possibly varying with n, provides motivation for this work.

To obtain risk bounds for PCR, we prove a master theorem, Theorem 17, that establishes a finite sample prediction risk bound for linear predictors of the general form

$$\widehat{Y}^*_{\widehat{B}} = X_*^\top \widehat{B} \left(\widehat{B}^\top \boldsymbol{X}^\top \boldsymbol{X} \widehat{B}\right)^+ \widehat{B}^\top \boldsymbol{X}^\top \boldsymbol{Y}, \tag{2.3}$$

where $\widehat{B} \in \mathbb{R}^{p \times q}$ is an appropriate matrix that may be deterministic or depend on the data $\boldsymbol{X}$, with dimension q allowed to be random.

This approach has the benefit of not only covering the special case of PCR,

corresponding to choice $\widehat{B} = \mathbf{U}_k$, but of offering a unifying analysis of other prediction schemes of the form (2.3). One important example corresponds to $\widehat{B} = \boldsymbol{I}_p$, which leads to another model agnostic predictor, the generalized least squares estimator (also known as the minimum norm interpolating predictor), which has enjoyed revamped popularity in the last two years [13–18, 33, 44, 51, 72, 77–79]. Using the full data matrix $\boldsymbol{X}$ for prediction—instead of just the first k principal components as in PCR—leads to additional bias compared to PCR prediction. However, in the high-dimensional regime $p \gg n$, this bias can become small and choosing $\widehat{B} = \boldsymbol{I}_p$ can become a viable alternative to PCR that requires no tuning parameters.

In addition to these two model-agnostic prediction methods, Theorem 17 can be used to analyze predictors directly tailored to model (2.1), which are shown formally to be of type (2.3) in Section 2.4.2. We give a particular expression of $\widehat{B}$, as well as the corresponding prediction analysis, under further modelling restrictions that render parameters K, A and β identifiable. The model specifications given in Section 2.4.2 allow us to view A as a cluster membership matrix, making it possible to address a third, understudied, class of examples pertaining to prediction from low-dimensional feature representation, that of prediction of Y via latent cluster centers, for features that exhibit an overlapping clustering structure corresponding to A.

2.1.1 Our Contributions and Organization of the Paper

Our main theoretical goal is to offer sufficient conditions on $\widehat{B}$ under which the prediction risk $\mathbb{R}(\widehat{B})$, defined as

$$\mathbb{R}(\widehat{B}) := \mathbb{E}[(Y_* - \widehat{Y}^*_{\widehat{B}})^2], \tag{2.4}$$

provably approaches an optimal risk benchmark, as n and p grow, with particular attention given to the case $p > n$. The expectation in (2.4) is taken with respect to the new data point (X_*, Y_*). Our main applications will be to the finite sample risk bounds of the three classes of predictors discussed in the previous section.

1. General finite sample risk bounds for linear predictors, under factor regression models. To meet our main theoretical goal, in Section 2.2, we state the risk benchmark in Lemma 16 and prove a master theorem, and our main theoretical result, Theorem 17. It provides a finite sample bound on $\mathbb{R}(\widehat{B})$, for generic $\widehat{B}$, when (X, Y) follow a factor regression model (2.1) that is fully introduced in Section 2.2.1.

The risk bound (2.14) of Theorem 17 depends on random quantities $\widehat{r} = \text{rank}(\boldsymbol{X}P_{\widehat{B}})$, $\widehat{\eta} = n^{-1}\sigma^2_{\widehat{r}}(\boldsymbol{X}P_{\widehat{B}})$, and $\widehat{\psi} = n^{-1}\sigma^2_1(\boldsymbol{X}P^{\perp}_{\widehat{B}})$, where we use $\sigma_k(M)$ to denote the kth largest singular value for any matrix M. To interpret these, note that $\widehat{Y}^*_{\widehat{B}} = \widehat{Y}^*_{P_{\widehat{B}}}$ (see Lemma 40 in Appendix B.2 for the proof), where $P_{\widehat{B}}$ is the projection onto the range of $\widehat{B}$. We then see that $\widehat{r}$ is the rank of the projected data matrix $\boldsymbol{X}P_{\widehat{B}}$ used for constructing $\widehat{Y}^*_{\widehat{B}}$, $\widehat{\eta}$ captures the size of the signal that is retained in $\boldsymbol{X}$ after projection onto the range of $\widehat{B}$, and $\widehat{\psi}$ captures the bias introduced by using only the component of $\boldsymbol{X}$ in the range of $\widehat{B}$ for prediction.

The utility of Theorem 17, as a general result, is in reducing the difficult task

of bounding $\mathbb{R}(\widehat{B})$ to the relatively easier one of controlling $\widehat{r}, \widehat{\eta}$, and $\widehat{\psi}$ corresponding to any matrix $\widehat{B}$ of interest.

2. Finite sample risk bounds for PCR-$\widehat{s}$, with data-adaptive $\widehat{s}$ principal components. We use Theorem 17 to analyze the prediction risk of PCR-$\widehat{s}$ under the factor regression model, for two choices of the number of principal components $\widehat{s}$. We first consider the *theoretical elbow method*, which selects $\widehat{s}$ corresponding to the smallest eigenvalue of $\boldsymbol{X}^\top \boldsymbol{X}/n$ above the noise level of order $\delta_W := c(\|\Sigma_W\|_{\text{op}} + \text{tr}(\Sigma_W)/n)$, for an absolute constant $c > 0$. Corollary 19 provides the rate

$$\mathbb{R}(\mathbf{U}_{\widehat{s}}) - \sigma^2 \lesssim (K + \log n)\frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1}\beta. \tag{2.5}$$

The first term on the right hand side is the standard variance term of linear regression in K dimensions. The second term is a bias term that arises from the fact that we predict using X instead of Z; we show that such a term is unavoidable in Lemma 16 of Section 2.2.2 below.

We termed this procedure *theoretical* as δ_W depends on unknown quantities of the data distribution. We address this by introducing a novel method in Section 2.3.1, which we show in Corollary 21 achieves the same rate as PCR with the theoretical elbow method, under mild additional assumptions, and is fully data-adaptive, only requiring the choice of one scale-free tuning parameter.

3. Minimum-norm interpolating predictors. In Section 2.4.1 we use the master theorem to recover risk bounds for the Generalized Least Squares predictor (GLS), independently derived in [33]. This predictor is also known as the minimum-norm interpolating predictor when $p > n$.

4. Prediction under identifiable factor regression models: Essential regression. In Section 2.4.2 we consider a particular identifiable factor regression model, the Essential Regression model introduced in [22]. The identifiability assumptions employ a type of errors-in-variables parametrization of A, described in Section 2.4.2, that allows the components of Z to be respectively matched with distinct groups of components of X. The latter property, combined with a further sparsity assumption on A, can be used to define overlapping clusters of X with latent centers Z_k, $1 \leq k \leq K$ [26]. Thus, of independent interest, prediction in Essential Regression is prediction via latent cluster centers. We show formally in Section 2.4.2 that this model specification leads to predictors of type (2.3), with $\widehat{B} = \widehat{A}$, for an appropriate estimator $\widehat{A}$ of A. We provide a finite sample prediction bound in Theorem 24, as an application of Theorem 17. We use the derived bound as an example that illustrates the possible benefits of sparsity in the predictor's coefficient matrix, as our matrix $\widehat{A}$ is allowed to be sparse.

5. Data-splitting under factor regression models. To allow for model selection among the diverse set of prediction methods in this setting, we offer a simple model selection approach in Section 2.5 based on data splitting. We provide an oracle inequality showing that the selected predictor performs nearly as well as the predictor with the lowest risk.

A preview of the results in Sections 2.3—2.4 is given in Table 2.1 below, which focuses on the high-dimensional regime where $p > Cn$ for a large enough constant $C > 0$, and is stated under the simplifying assumptions $\lambda_K(A^\top A) \gtrsim p/K$

and $r_e(\Sigma_W) \asymp p$, where $r_e(\Sigma_W) := \mathrm{tr}(\Sigma_W)/\|\Sigma_W\|_{\mathrm{op}}$ is the reduced effective rank of Σ_W, the covariance matrix of W from model (2.1). The bound for Essential Regression contains the quantity $\|A_J\|_0$, which is the sparsity level of the sub-matrix A_J of A corresponding to *non-pure* variables in the Essential Regression model, namely the variables associated with more than one latent factor Z_k (see Section 2.4.2 for a formal definition). The full set of conditions under which these bounds hold, as well as their general form is given, respectively, in each of the sections in which these methods are analyzed. For now we mention that we do not make specific distributional assumption on the data, but we do derive the rates given in the table below under the assumption that $\varepsilon \in \mathbb{R}$, $Z \in \mathbb{R}^K$, and $W \in \mathbb{R}^p$ are sub-Gaussian.

The term $\sigma^2 K/n$ is common to all three risk bounds, and shows that all methods have the potential to adapt to the unknown, latent, K-dimensional model structure, provided that the remaining terms are small. Relative to PCR and ER, the GLS method has an additional variance term $\sigma^2 n/p$, that arises from the fact that GLS uses the full data matrix $\boldsymbol{X}$, as opposed to a lower-dimensional projection of it; this demonstrates that GLS has competitive performance only when $p \gg n$. The relative performance of the PCR and ER methods depends on the sparsity of the matrix A_J: when $\|A_J\|_0 = o(p)$, for example, the ER method can outperform PCR.

We further discuss the relative merits of these predictors, in terms of their respective risk bounds and assumptions under which they hold, in Section 2.4.3.

We conclude the paper with Section 2.6, in which we present a detailed simulation study of the PCR-type predictors, the minimum-norm interpolating predictor, and predictors under Essential Regression, as well as the proposed

Prediction Method	$\widehat{B}$	Excess risk bound
PCR	$\mathbf{U}_K$	$\frac{K}{n}\sigma^2 + \frac{K}{p}\|\Sigma_W\|_{\text{op}}\|\beta\|^2 + \frac{K}{n}\|\Sigma_W\|_{\text{op}}\|\beta\|^2$
GLS	$\boldsymbol{I}_p$	$\frac{K}{n}\sigma^2 + \frac{n}{p}\sigma^2 + \frac{K}{n}\|\Sigma_W\|_{\text{op}}\|\beta\|^2$
ER	$\widehat{A}$	$\frac{K}{n}\sigma^2 + \frac{K}{p}\|\Sigma_W\|_{\text{op}}\|\beta\|^2 + \frac{\|A_J\|_0}{p} \times \frac{K}{n}\|\Sigma_W\|_{\text{op}}\|\beta\|^2$

Table 2.1: Summary of bounds on $\mathbb{R}(\widehat{B}) - \sigma^2$, where $\mathbb{R}(\widehat{B})$ is defined in (2.4), for Principal Component Regression (PCR), Generalized Least Squares (GLS), and Essential Regression (ER), stated under simplifying assumptions described in Section 2.4.3. The second column gives the choice of $\widehat{B}$ corresponding to each method. All three bounds follow from the main Theorem 17.

model selection method. All proofs are deferred to the Appendix.

Notation: We use the following notation throughout the paper. For any vector v, we use $\|v\|_q$ denote its ℓ_q norm for $0 \le q \le \infty$. We write $\|v\| = \|v\|_2$. For an arbitrary real-valued matrix $M \in \mathbb{R}^{r\times q}$, we use M^+ to denote the Moore-Penrose inverse of M, and $\sigma_1(M) \ge \sigma_2(M) \ge \cdots \ge \sigma_{\min(r,q)}(M)$ to denote the singular values of M in non-increasing order. We define the operator norm $\|M\|_{\text{op}} = \sigma_1(M)$, the Frobenius norm $\|M\|_F^2 = \sum_{i,j} M_{ij}^2$, the elementwise sup-norm $\|M\|_\infty = \max_{i,j}|M_{ij}|$ and the cardinality of non-zero entries $\|M\|_0 = \sum_{i,j} 1_{M_{ij}\neq 0}$. For a symmetric positive semi-definite matrix $Q \in \mathbb{R}^{p\times p}$, we use $\lambda_1(Q) \ge \lambda_2(Q) \ge \cdots \ge \lambda_p(Q)$ to denote the eigenvalues of Q in non-increasing order, and $\kappa(Q) = \lambda_1(Q)/\lambda_p(Q)$ to denote its condition number.

For any two sequences a_n and b_n, we write $a_n \lesssim b_n$ if there exists some constant C such that $a_n \le Cb_n$. The notation $a_n \asymp b_n$ stands for $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

We use $\boldsymbol{I}_d$ to denote the $d \times d$ identity matrix. For $m \ge 1$, we let $[m] = \{1, 2, \ldots, m\}$. Lastly, we use c, c', C, C' to denote positive and finite absolute con-

stants that unless otherwise indicated can change from line to line.

2.2 Bounding the Risk $\mathbb{R}(\widehat{B})$

In this section we derive and discuss bounds on the risk $\mathbb{R}(\widehat{B})$ defined in (2.4), corresponding to the predictor $\widehat{Y}^*_{\widehat{B}}$. Our results are valid for any $\widehat{B} \in \mathbb{R}^{p\times q}$ that can be either random depending on $\boldsymbol{X}$ or fixed, where $q \leq p$ but is allowed to be random.

2.2.1 Preliminaries

As the risk $\mathbb{R}(\widehat{B})$ is defined relative to the first two moments of (X, Y), which are further linked to quantities $(A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ under model (2.1), our risk bounds are written in terms of the components of $\theta := (K, \beta, A, \Sigma_Z, \Sigma_W, \sigma^2)$. We thus start by formally defining model (2.1) with respect to θ.

[(Sub-Gaussian) Factor Regression Model] We say the pair (X, Y) follows the model FRM(θ) with $\theta = (K, \beta, A, \Sigma_Z, \Sigma_W, \sigma^2)$, and write $(X, Y) \sim \mathbb{P}_\theta$ or $(X, Y) \sim$ FRM(θ), when

(1) Equation (2.1) holds with matrix $A \in \mathbb{R}^{p\times K}$, vector $\beta \in \mathbb{R}^K$, and random quantities $(Z, W, \varepsilon) \in (\mathbb{R}^K, \mathbb{R}^p, \mathbb{R})$ that are mutually independent;

(2) W and ε are mean zero with $\mathbb{E}_\theta[WW^\top] = \Sigma_W$ and $\mathbb{E}_\theta[\varepsilon^2] = \sigma^2$, and Z is also mean zero without loss of generality, with $\mathbb{E}_\theta[ZZ^\top] = \Sigma_Z$.

(3) Both A and Σ_Z have rank equal to K.

We further say $(X, Y) \sim$ sG-FRM(θ) if the following holds in addition to (1)—(3)

(4) There exist finite, absolute positive constants $\gamma_\varepsilon, \gamma_w$ and γ_z such that

(a) ε is $\sigma\gamma_\varepsilon$ sub-Gaussian;[1]

(b) $Z = \Sigma_Z^{1/2}\widetilde{Z}$ where $\widetilde{Z}$ is γ_z sub-Gaussian with $\mathbb{E}_\theta[\widetilde{Z}\widetilde{Z}^\top] = \boldsymbol{I}_K$;[2]

(c) $W = \Sigma_W^{1/2}\widetilde{W}$ where $\widetilde{W}$ is γ_w sub-Gaussian with $\mathbb{E}_\theta[\widetilde{W}\widetilde{W}^\top] = \boldsymbol{I}_p$.

Since there exist multiple parameters θ for which (X, Y) has the same joint distribution, the model is not identifiable without further restrictions on the parameter space. As this work is devoted to the prediction of Y, and not to the estimation of θ, this is not problematic. We thus allow for this lack of identifiability and our subsequent analysis of $\mathbb{R}(\widehat{B}) := \mathbb{E}_\theta[(Y_* - \widehat{Y}^*_{\widehat{B}})^2]$ is valid for any θ such that $(X, Y) \sim$ sG-FRM(θ). In particular, the analysis is applicable to any identifiable sG-FRM(θ), whenever further structure on θ is added to Definition 2.2.1. We note that $\mathbb{R}(\widehat{B})$ depends on θ, but we suppress this dependence in the notation for simplicity.

2.2.2 Benchmark of $\mathbb{R}(\widehat{B})$

To provide a benchmark for $\mathbb{R}(\widehat{B})$, we let

$$\alpha^* := \arg\min_\alpha \mathbb{E}\left[(Y_* - X_*^\top\alpha)^2\right] = [\mathrm{Cov}(X)]^+\mathrm{Cov}(X, Y) \tag{2.6}$$

denote the coefficient of the best linear predictor (BLP) of Y_* from X_*, where $[\mathrm{Cov}(X)]^+$ is the Moore-Penrose pseudoinverse of $\mathrm{Cov}(X)$. For any $\theta =$

[1] A mean zero random variable x is called γ sub-Gaussian if $\mathbb{E}[\exp(tx)] \le \exp(t^2\gamma^2/2)$ for all $t \in \mathbb{R}$.

[2] A mean zero random vector x is called γ sub-Gaussian if $\langle x, v\rangle$ is γ sub-Gaussian for any unit vector v.

$(K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ such that $(X_*, Y_*) \sim \text{FRM}(\theta)$ with corresponding latent vector Z_*, we have the following chain of simple equalities from our independence assumptions

$$\begin{aligned}
\mathbb{R}(\widehat{B}) &= \mathbb{E}_\theta\left[(Y_* - X_*^\top \alpha^*)^2\right] + \mathbb{E}_\theta\left[(X_*^\top \alpha^* - \widehat{Y}^*_{\widehat{B}})^2\right] \\
&= \sigma^2 + \mathbb{E}_\theta\left[(Z_*^\top \beta - X_*^\top \alpha^*)^2\right] + \mathbb{E}_\theta\left[(X_*^\top \alpha^* - \widehat{Y}^*_{\widehat{B}})^2\right] \qquad (2.7) \\
&= \sigma^2 + \mathbb{E}_\theta\left[(Z_*^\top \beta - \widehat{Y}^*_{\widehat{B}})^2\right].
\end{aligned}$$

We interpret the term $\sigma^2 = \mathbb{E}_\theta[\varepsilon^2]$ as an oracle risk value because it is the minimal risk of predicting Y_* from Z_*, had Z_* been observable. We thus focus on bounding the difference $\mathbb{R}(\widehat{B}) - \sigma^2$ and refer to it as *excess risk*, with the tacit understanding that the excess is relative to oracle prediction.

We further note that the term $\mathbb{E}_\theta[(Z_*^\top \beta - X_*^\top \alpha^*)^2]$ in (2.7) is the minimal risk incurred by predicting $Z_*^\top \beta$ by $X_*^\top \alpha^*$, with an observable X_*. Display (2.7) shows that it is a population level cost that is incurred in any risk analysis of a predictor of type (2.3) performed under FRM(θ). Lemma 16 below quantifies its size, and makes use of the signal-to-noise ratio given by

$$\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_W\|_{\text{op}}. \qquad (2.8)$$

Its proof can be found in Appendix B.2.1.

Lemma 16. *For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with invertible Σ_W such that $(X, Y) \sim$ FRM(θ),*

$$\frac{\xi}{1+\xi}\beta^\top (A^\top \Sigma_W^{-1} A)^{-1}\beta \le \mathbb{E}_\theta\left[(Z_*^\top \beta - X_*^\top \alpha^*)^2\right] \le \beta^\top (A^\top \Sigma_W^{-1} A)^{-1}\beta. \qquad (2.9)$$

The inequalities above become asymptotically tight when the signal retained in K dimensions by X dominates the ambient noise, that is, when $\xi \to \infty$ as $p \to \infty$. In general, as soon as $\xi > c$, for some $c > 0$ and Σ_W is well conditioned

such that $\kappa(\Sigma_W) = \lambda_1(\Sigma_W)/\lambda_p(\Sigma_W) < C$, we further obtain, using (2.7), for any $\widehat{B}$, that

$$\mathbb{R}(\widehat{B}) - \sigma^2 \geq \mathbb{E}_\theta\left[(Z_*^\top\beta - X_*^\top\alpha^*)^2\right] \gtrsim \|\Sigma_W\|_{\mathrm{op}}\beta^\top(A^\top A)^{-1}\beta. \tag{2.10}$$

Therefore a risk analysis of linear predictors under factor regression models, which consists in upper bounding $\mathbb{R}(\widehat{B}) - \sigma^2$, will necessarily include terms larger than $\|\Sigma_W\|_{\mathrm{op}}\beta^\top(A^\top A)^{-1}\beta$ in the risk bounds, irrespective of the construction of the linear predictor. If, in addition, $A\Sigma_Z A^\top$ is well-conditioned with $\lambda_1(A\Sigma_Z A^\top)/\lambda_K(A\Sigma_Z A^\top) \leq C$, then

$$\beta^\top(A^\top\Sigma_W^{-1}A)^{-1}\beta \asymp \|\Sigma_W\|_{\mathrm{op}}\beta^\top\Sigma_Z^{1/2}\left(\Sigma_Z^{1/2}A^\top A\Sigma_Z^{1/2}\right)^{-1}\Sigma_Z^{1/2}\beta \asymp \frac{\beta^\top\Sigma_Z\beta}{\xi}$$

and Lemma 16 in turn implies

$$\frac{\beta^\top\Sigma_Z\beta}{1+\xi} \lesssim \mathbb{E}_\theta\left[(Z_*^\top\beta - X_*^\top\alpha^*)^2\right] \lesssim \frac{\beta^\top\Sigma_Z\beta}{\xi}.$$

This demonstrates that the signal-to-noise ratio ξ must necessarily dominate $\beta^\top\Sigma_Z\beta$ for the excess risk $\mathbb{R}(\widehat{B}) - \sigma^2$ to vanish as $p \to \infty$.

2.2.3 Upper Bound of the Risk $\mathbb{R}(\widehat{B})$

To motivate our main result, we first introduce some key quantities that appear in the risk bound derivation for any generic $\widehat{B}$ leading to the predictors of type (2.3).

The prediction risk bound depends on W in Definition 2.2.1, specifically on the noise level of $n^{-1}\|\boldsymbol{W}^\top\boldsymbol{W}\|_{\mathrm{op}}$. To quantify this noise level, we use the following deviation bound from Lemma 47 in Appendix B.3. For any θ such that $(X, Y) \sim$

sG-FRM(θ), one has

$$\mathbb{P}_\theta \left\{ \frac{1}{n} \|\boldsymbol{W}^\top \boldsymbol{W}\|_{\mathrm{op}} \le \delta_W \right\} \ge 1 - e^{-n} \tag{2.11}$$

where δ_W is defined as

$$\delta_W := \delta_W(\theta) = c \left[\|\Sigma_W(\theta)\|_{\mathrm{op}} + \frac{\mathrm{tr}(\Sigma_W(\theta))}{n} \right], \tag{2.12}$$

with $c = c(\gamma_w)$ being some positive constant. The quantity δ_W will play a role in the risk bound and it could take any non-negative value in general. When $\lambda_1(\Sigma_W) \le C$ for some constant $C > 0$, one has $\delta_W \lesssim 1 + p/n$. When $\lambda_p(\Sigma_W) \ge c$ for some constant $c > 0$, we have $\delta_W \gtrsim 1 + p/n$. In particular, if $c \le \lambda_p(\Sigma_W) \le \lambda_1(\Sigma_W) \le C$, we have $\delta_W \asymp 1 + p/n$. This holds for instance when Σ_W is diagonal with entries bounded away from 0 and ∞, independent of n.

We write the projection onto the column space of $\widehat{B}$ as

$$P_{\widehat{B}} = \widehat{B}[\widehat{B}^\top \widehat{B}]^+ \widehat{B}^\top = \widehat{B}\widehat{B}^+,$$

its complement as $P^\perp_{\widehat{B}} = \boldsymbol{I}_p - P_{\widehat{B}}$ and $\widehat{r} = \mathrm{rank}(\boldsymbol{X}P_{\widehat{B}})$. Since $\widehat{B}[\boldsymbol{X}\widehat{B}]^+ = P_{\widehat{B}}[\boldsymbol{X}P_{\widehat{B}}]^+$, as proved in Lemma 40 in Appendix B.2, we find that $\widehat{Y}^*_{\widehat{B}} = X_*^\top \widehat{B}[\boldsymbol{X}\widehat{B}]^+ \boldsymbol{Y} = \widehat{Y}^*_{P_{\widehat{B}}}$ making clear that the component of the data matrix orthogonal to the range of $\widehat{B}$, $\boldsymbol{X}P^\perp_{\widehat{B}}$, is not used for prediction. It is natural therefore that the size of this component, as measured by its largest singular value, $\sigma_1^2(\boldsymbol{X}P^\perp_{\widehat{B}})$, will affect the risk bound, and needs to be contrasted with the size of the retained signal, $\boldsymbol{X}P_{\widehat{B}}$, as measured by its smallest non-zero singular value $\sigma^2_{\widehat{r}}(\boldsymbol{X}P_{\widehat{B}})$. These two quantities appear in the risk bound below.

We now state our main theorem; its proof is deferred to Appendix B.2.1. Recall that $\mathbb{R}(\widehat{B})$ is the risk defined in (2.4). Write $a \wedge b = \min\{a, b\}$.

Theorem 17. *Let $\widehat{B} = \widehat{B}(\boldsymbol{X}) \in \mathbb{R}^{p\times q}$ for some $q \ge 1$, and set*

$$\widehat{r} := \mathit{rank}\left(\boldsymbol{X}P_{\widehat{B}}\right), \qquad \widehat{\eta} := \frac{1}{n}\sigma^2_{\widehat{r}}\left(\boldsymbol{X}P_{\widehat{B}}\right), \qquad \widehat{\psi} := \frac{1}{n}\sigma_1^2\left(\boldsymbol{X}P^\perp_{\widehat{B}}\right). \tag{2.13}$$

For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \le Cn/\log n$ for some positive constant $C = C(\gamma_z)$ such that $(X, Y) \sim$ sG-FRM(θ), there exists some absolute constant $c > 0$ such that

$$\mathbb{P}_\theta\Big\{\mathbb{R}(\widehat{B}) - \sigma^2 \lesssim \left[\frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\eta}}\widehat{r} + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)(K \wedge \widehat{r} + \log n)\right]\frac{\sigma^2}{n} \\ + \left[\left(1 + \frac{\|\Sigma_W\|_{op}}{\widehat{\eta}}\right)\delta_W + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)\widehat{\psi}\right]\beta^\top(A^\top A)^{-1}\beta\Big\} \ge 1 - c/n. \tag{2.14}$$

Here the symbol $\lesssim$ means the inequality holds up to a multiplicative constant possibly depending on the sub-Gaussian constants γ_ε, γ_z and γ_w.

Since we aim to provide a unified analysis of the risk for a general $\widehat{B}$, the bound (2.14) itself depends on the random quantities $\widehat{r}$, $\widehat{\eta}$ and $\widehat{\psi}$. To make it informative, one needs to further control these random quantities for specific choices of $\widehat{B}$. The main usage of Theorem 17 is thus to reduce the task of bounding $\mathbb{R}(\widehat{B})$ to the relatively easier one of controlling $\widehat{r}$, $\widehat{\eta}$ and $\widehat{\psi}$. We will demonstrate this for several choices of $\widehat{B}$ in the following sections.

Theorem 17 holds for any estimator $\widehat{B} \in \mathbb{R}^{p\times q}$ that is constructed from $\boldsymbol{X}$ with any $q \ge 1$. We now explain the various terms in the bound (2.14). Recall that $\widehat{Y}^*_{\widehat{B}} = X_*^\top \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Y}$ and $\boldsymbol{Y} = \boldsymbol{Z}\beta + \boldsymbol{\varepsilon}$. To aid intuition, by adding and subtracting terms, we have

$$\begin{aligned}\widehat{Y}^*_{\widehat{B}} - Z_*^\top\beta &= X_*^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon} + X_*^\top\alpha^* - Z_*^\top\beta + X_*^\top\left[\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta - \alpha^*\right] \\ &= X_*^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon} + (X_*^\top\alpha^* - Z_*^\top\beta) + X_*^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{Z}\beta - \boldsymbol{X}\alpha^*) \\ &\quad + X_*^\top\left[\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X} - \boldsymbol{I}_p\right]\alpha^*. \end{aligned} \tag{2.15}$$

We discuss the four terms above one by one.

- The first term leads to the following variance term in (2.14):

$$\left[\frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\eta}}\widehat{r} + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)(K \wedge \widehat{r} + \log n)\right]\frac{\sigma^2}{n}.$$

We see that the random variable $\widehat{\eta}$ quantifies the retained signal in $\widehat{B}(\boldsymbol{X}\widehat{B})^+$ by noting that $\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\|_{\mathrm{op}}^2 = \|P_{\widehat{B}}(\boldsymbol{X}P_{\widehat{B}})^+\|_{\mathrm{op}}^2 \leq (n\widehat{\eta})^{-1}$. The two factors $\|\Sigma_W\|_{\mathrm{op}}/\widehat{\eta}$ and $(1 + \delta_W/\widehat{\eta})$ come from bounding the second moments of W_* and AZ_* from $X_* = AZ_* + W_*$, respectively, relative to the retained signal $\widehat{\eta}$. The dimension $\widehat{r}$ reflects the complexity of $\boldsymbol{X}P_{\widehat{B}}$ and the integer K is the intrinsic dimension of the latent factor, thus only appearing in the term containing $(1 + \delta_W/\widehat{\eta})$.

- The second and third terms in (2.15) lead to the following term in (2.14), which can be interpreted as arising from the fact that Z_* and $\boldsymbol{Z}$ are not observed:

$$\left(1 + \frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\eta}}\right)\frac{\delta_W}{\|\Sigma_W\|_{\mathrm{op}}} \cdot \|\Sigma_W\|_{\mathrm{op}}\beta^\top(A^\top A)^{-1}\beta.$$

  With slight abuse of terminology, we refer to this as a bias term. The factor

$$\|\Sigma_W\|_{\mathrm{op}}\beta^\top(A^\top A)^{-1}\beta$$

  is irreducible, as argued in (2.10), the term $\|\Sigma_W\|_{\mathrm{op}}/\widehat{\eta}$ has been explained in the first term, and the inflation factor $\delta_W/\|\Sigma_W\|_{\mathrm{op}}$ is due to the inflated noise level of $n^{-1}\|\boldsymbol{W}^\top\boldsymbol{W}\|_{\mathrm{op}}$ compared to $\|\Sigma_W\|_{\mathrm{op}}$.

- The fourth term in (2.15) quantifies the error of estimating the best linear predictor α^* under the factor regression model. In this model, we note that $\alpha^* = \Sigma^+A\Sigma_Z\beta$ with $\Sigma := \mathrm{Cov}(X)$. Also noting that $\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}$ is a projection matrix, the fourth term in (2.15) represents the error of estimating the range space of Σ^+A, which is exactly zero if the range of $\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}$ contains the range of Σ^+A. In general, the bound in (2.14) corresponding to this term is

$$\delta_W\beta^\top(A^\top A)^{-1}\beta + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)\widehat{\psi} \cdot \beta^\top(A^\top A)^{-1}\beta,$$

  where the first part is the error of estimating the range space of $P_{\widehat{B}}\Sigma^+A$ while

the second part is that of estimating the range space of $P_{\widehat{B}}^{\perp}\Sigma^{+}A$, controlled by $\widehat{\psi}$.

Remark 5. In light of the above discussion, we make two important remarks. First, to maintain a fast rate of the risk bound in (2.14), we should retain enough signal in $\boldsymbol{X}P_{\widehat{B}}$ relative to the noise δ_W such that $\widehat{\eta} \gtrsim \delta_W$ with high probability. Second, if this is the case, the bound (2.14) simplifies to

$$\mathbb{R}(\widehat{B}) - \sigma^2 \lesssim \left[\frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\eta}}\widehat{r} + (K \wedge \widehat{r} + \log n)\right]\frac{\sigma^2}{n} + \left(\delta_W + \widehat{\psi}\right)\beta^{\top}(A^{\top}A)^{-1}\beta.$$

As $\widehat{r} = \mathrm{rank}(\boldsymbol{X}P_{\widehat{B}})$ increases, meaning that the predictor can be interpreted as more complex, the variance term increases, while the term $\delta_W\beta^{\top}(A^{\top}A)^{-1}\beta$ is not affected.

If $\widehat{\psi}$ decreases as $\widehat{r}$ increases (as seen with the PCR predictor studied in the next section), the term $\widehat{\psi}\beta^{\top}(A^{\top}A)^{-1}\beta$, corresponding to the error of estimating the range space of $P_{\widehat{B}}^{\perp}\Sigma^{+}A$, gets smaller.

Therefore, the tradeoff of using a more complex predictor lies between the increasing variance and the decreasing error of estimating the range space of $P_{\widehat{B}}^{\perp}\Sigma^{+}A$, provided that enough signal is retained in $\boldsymbol{X}P_{\widehat{B}}$. A more transparent tradeoff can be seen for the PCR predictor analyzed in the next section. More generally, for each of our examples, we will see the mechanism by which $\widehat{r}$, $\widehat{\eta}$, and $\widehat{\psi}$ are controlled.

2.3 Analysis of Principal Component Regression Under the Factor Regression Model

In this section we use the general result, Theorem 17, to derive risk bounds for the popular Principal Component Regression (PCR) method. For any integer $1 \le k \le \text{rank}(\boldsymbol{X})$, the PCR-predictor PCR-$k$ corresponds to taking $\widehat{B} = \mathbf{U}_k$, the $p \times k$ matrix with columns equal to the first k right singular vectors of $\boldsymbol{X}$ corresponding to the non-increasing singular values $\sigma_1(\boldsymbol{X}) \ge \sigma_2(\boldsymbol{X}) \ge \cdots$. We start by giving risk bounds for PCR-k for any k in the corollary below. For simplicity, we write

$$\widehat{\lambda}_k = \frac{1}{n}\sigma_k^2(\boldsymbol{X})$$

with the convention that $\widehat{\lambda}_0 = \infty$ and $\widehat{\lambda}_k = 0$ for all $k > \text{rank}(\boldsymbol{X})$. All the proofs of this section can be found in Appendix B.2.2.

Corollary 18. *For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \le Cn/\log n$ and some positive constant $C = C(\gamma_z)$ such that (X, Y) follows sG-FRM(θ), there exists some absolute constant $c > 0$ such that, for any k (possibly random),*

$$\mathbb{P}_\theta\left\{\mathbb{R}(\mathbf{U}_k) - \sigma^2 \lesssim \widehat{B}(k)\right\} \ge 1 - cn^{-1} \tag{2.16}$$

where $\widehat{B}(k) = \widehat{B}_1(k) + \widehat{B}_2(k)$ and

$$\widehat{B}_1(k) := \left[\frac{\|\Sigma_W\|_{\text{op}}}{\widehat{\lambda}_k}k + \left(1 + \frac{\delta_W}{\widehat{\lambda}_k}\right)(K \wedge k + \log n)\right]\frac{\sigma^2}{n} \tag{2.17}$$

$$\widehat{B}_2(k) := \left(\frac{\|\Sigma_W\|_{op}}{\widehat{\lambda}_k}\delta_W + \delta_W + \widehat{\lambda}_{k+1}\right)\beta^\top(A^\top A)^{-1}\beta. \tag{2.18}$$

Corollary 18 follows immediately from the identities $\sigma_k^2(\boldsymbol{X}P_{\mathbf{U}_k}) = \sigma_k^2(\boldsymbol{X})$ and $\sigma_1^2(\boldsymbol{X}P^\perp_{\mathbf{U}_k}) = \sigma^2_{k+1}(\boldsymbol{X})$, and an application of Theorem 17 with

$$\widehat{r} = k, \qquad \widehat{\eta} = \widehat{\lambda}_k, \qquad \widehat{\psi} = \widehat{\lambda}_{k+1} \qquad \text{almost surely.}$$

The bound $\widehat{B}(k)$ in (2.16) depends on $\widehat{\lambda}_k$ and $\widehat{\lambda}_{k+1}$, which may be further controlled by $\lambda_k(A\Sigma_Z A^\top) - \delta_W$ and $\lambda_{k+1}(A\Sigma_Z A^\top) + \delta_W$, respectively, in order to make the bound more informative (see, for example, the proof of Remark 6 in Appendix B.2.2). Nevertheless, (2.16) illustrates the effect of k and hints at the choice $k = \widehat{s}$ with

$$\widehat{s} = \max\left\{k \geq 0 : \ \widehat{\lambda}_k \geq C_0 \delta_W\right\}. \tag{2.19}$$

Here δ_W is defined in (2.12) and C_0 is some positive constant. The quantity $\widehat{s}$ corresponds to what is known as the *elbow method*, and is a ubiquitous approach for selecting the number of top principal components of the data matrix $\boldsymbol{X}$. The quality of $\widehat{s}$ as an estimator of the effective rank of $\Sigma = \mathrm{Cov}(X)$ has been analyzed in [34], but its role in PCR has received little attention. By definition, $\widehat{\lambda}_{\widehat{s}+1} < C_0\delta_W \leq \widehat{\lambda}_{\widehat{s}}$, which implies

$$\widehat{B}(\widehat{s}) \lesssim (\widehat{s} + \log n)\frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1}\beta, \qquad \text{almost surely.}$$

Furthermore, Weyl's inequality implies $\widehat{\lambda}_{K+1} \leq \sigma_1^2(\boldsymbol{W})/n$ and, in conjunction with (2.11), and by choosing $C_0 > 1$, we obtain $\widehat{s} \leq K$ with high probability. We summarize this discussion in the following result pertaining to prediction via the first $\widehat{s}$ principal components selected via the elbow method.

Corollary 19. *For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \leq Cn/\log n$ such that (X, Y) follows sG-FRM(θ), we have for $\widehat{s}$ defined in (2.19) for any $C_0 > 1$,*

$$\mathbb{P}_\theta\left\{\mathbb{R}(\mathbf{U}_{\widehat{s}}) - \sigma^2 \lesssim (K + \log n)\,\frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1}\beta\right\} \geq 1 - O(n^{-1}). \tag{2.20}$$

Remark 6.

1. We refer to the method analyzed in Corollary 19 as the *theoretical* elbow method, as it involves the theoretically optimal threshold level δ_W. The next section analyzes the performance of a *data-adaptive* elbow method.

2. For any θ, we show in Appendix B.2.2 that, if $\lambda_K(A\Sigma_Z A^\top) \geq C\delta_W$ for some sufficiently large constant $C > 0$, then $\widehat{\lambda}_K \geq C_0\delta_W$ holds for some $C_0 > 1$ with high probability. The event $\{\widehat{\lambda}_K \geq C_0\delta_W\}$ implies $\{\widehat{s} \geq K\}$ which, in conjunction with the high probability event $\{\widehat{s} \leq K\}$, guarantees $\widehat{s} = K$ with high probability. Corollary 19 thus covers the risk of PCR-K, that is, the risk of the PCR predictor corresponding to the true K of this θ.

2.3.1 Selection of the Number of Retained Principal Components via Penalized Least Squares

A practical issue of PCR-$\widehat{s}$ is that the selection of $\widehat{s}$ according to (2.19) relies on a theoretical order δ_W in (2.12), which depends on the unknown quantities $\|\Sigma_W\|_{\mathrm{op}}$ and $\mathrm{tr}(\Sigma_W)$. To overcome this difficulty, we provide an alternative, data dependent procedure, which shares the risk bound derived for PCR-$\widehat{s}$.

Our procedure of selecting the number of retained principal components is adopted from [27], originally proposed for selecting the rank of the coefficient of a multivariate response regression model $\boldsymbol{Y} = \boldsymbol{X}B + \boldsymbol{W}$. The factor model $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$ is a particular case with $\boldsymbol{X} = \boldsymbol{I}_{n\times p}$ and $B = \boldsymbol{Z}A^\top$, and, following [27], we define

$$\widetilde{s} := \operatorname*{arg\,min}_{0\leq k\leq \bar{K}} \widehat{v}_k^2, \quad \text{with} \quad \widehat{v}_k^2 := \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(k)}\|_F^2}{np - \mu_n k}, \quad \text{and} \quad \bar{K} := \left\lfloor \frac{\kappa}{1+\kappa}\frac{np}{\mu_n} \right\rfloor \wedge n \wedge p, \tag{2.21}$$

for a given sequence $\mu_n > 0$. Here $\kappa > 1$ is some absolute constant introduced to avoid division by zero. We write $\boldsymbol{X}_{(k)}$ as the best rank k approximation of $\boldsymbol{X}$. More specifically, let the SVD of $\boldsymbol{X}$ as $\boldsymbol{X} = \sum_j \sigma_j u_j v_j^\top$ with non-increasing σ_j and

we have $\boldsymbol{X}_{(k)} = \sum_{j=1}^{k} \sigma_j u_j v_j^\top$.

The denominator of the ratio defining $\widehat{v}_k^2$ can be viewed as a penalty on the numerator, with tuning sequence μ_n. From [27, Equation 2.7], the minimizer $\widetilde{s}$ conveniently has a closed form

$$\widetilde{s} = \sum_k 1\{\widehat{\lambda}_k \geq \mu_n \widehat{v}_k^2\},$$

counting the number of singular values of $\boldsymbol{X}$ above a *variable* threshold. This is in contrast to the elbow method in (2.19), which counts the number of singular values of $\boldsymbol{X}$ above the *fixed* threshold $\mu = C_0 \delta_W$, as

$$\widehat{s} = \sum_k 1\{\widehat{\lambda}_k \geq \mu\}.$$

We note that when $\Sigma_W = 0$, $\|\boldsymbol{X} - \boldsymbol{X}_{(k)}\|_F = \|\boldsymbol{Z}A^\top - (\boldsymbol{Z}A^\top)_{(k)}\|_F = 0$ for any $k \geq K$. Hence there are multiple minima (zeroes in this case) in $\widehat{v}_k^2$, and if we adopt the convention to choose the first index k with $\|\boldsymbol{X} - \boldsymbol{X}_{(k)}\|_F = 0$, we find $\widetilde{s} = K$, almost surely. The risk of PCR-K has already been discussed in Remark 6 above.

The theoretical guarantees proved in [27] are based on the assumption that $\boldsymbol{W}$ has i.i.d. entries with zero mean and bounded fourth moments. Proposition 20 extends this to models in which the rows of $\boldsymbol{W}$ are allowed to have dependent entries, when they follow a sub-Gaussian distribution. We show that the choice $\mu_n = c_0(n + p)$, for some absolute numerical constant c_0, leads to desirable results. The induced size of $\bar{K}$, for this μ_n, is of order $n \wedge p$. We found the choice $c_0 = 0.25$ worked well for all our simulations, as presented in Section 2.6.

Let $r_e(\Sigma_W) = \mathrm{tr}(\Sigma_W)/\|\Sigma_W\|_{\mathrm{op}}$ denote the effective rank of Σ_W. The following proposition shows that $\widetilde{s}$ finds, adaptively, the *theoretical* elbow.

Proposition 20. *Let $\widetilde{s}$ be defined in (2.21) with $\mu_n = c_0(n+p)$ for some absolute constant $c_0 > 0$. For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ such that (X, Y) follows sG-FRM(θ), $\log p \le cn$, $K \le \bar{K}$ and*

$$r_e(\Sigma_W) \ge c'(n \wedge p) \tag{2.22}$$

for some positive constants $c = c(\gamma_w)$ and $c' = c'(\gamma_w)$, we have

$$\mathbb{P}_\theta \left\{ \widetilde{s} \le K, \quad \widehat{\lambda}_{\widetilde{s}} \gtrsim \delta_W, \quad \widehat{\lambda}_{\widetilde{s}+1} \lesssim \delta_W \right\} \ge 1 - O(1/n). \tag{2.23}$$

Condition $K \le \bar{K}$ holds, for instance, if $K \le c''(n \wedge p)$ with $c'' \le \kappa/(2c_0(1+\kappa))$. We explain the connection between restriction (2.22) and the proposed choice of μ_n. Using elementary algebra, [27, Theorem 6 and Proposition 7] proves the deterministic result

$$\left\{ \frac{2\sigma_1^2(\boldsymbol{W})}{\|\boldsymbol{W}\|_F^2/(np)} \le \mu_n \right\} \subseteq \{\widetilde{s} \le K\}, \tag{2.24}$$

which shows that if μ_n is appropriately large, then the selected $\widetilde{s}$ is less than or equal to dimension K of the factor regression model generating the data. On the other hand, by concentration inequalities of $\|\boldsymbol{W}\|_F^2/n$ and $\sigma_1^2(\boldsymbol{W})/n$ around $\mathrm{tr}(\Sigma_W)$ and δ_W, respectively (see the proof of Proposition 20 in Appendix B.2.2), the bound

$$\frac{2\sigma_1^2(\boldsymbol{W})}{\|\boldsymbol{W}\|_F^2/(np)} \lesssim np\frac{\delta_W}{\mathrm{tr}(\Sigma_W)} = p + \frac{np}{r_e(\Sigma_W)} \tag{2.25}$$

holds with probability larger than $1 - O(1/n)$. Thus, in view of (2.24) and (2.25), the event $\{\widetilde{s} \le K\}$ holds with high probability as soon as $\mu_n > p + np/r_e(\Sigma_W)$. Under (2.22), we arrive at the choice $\mu_n = c_0(n+p)$ and, in turn, $\bar{K} = O(n \wedge p)$.

We note that (2.22) holds, for instance, in the commonly considered setting

$$0 < c' \le \lambda_p(\Sigma_W) \le \lambda_1(\Sigma_W) \le C' < \infty, \tag{2.26}$$

while being more general. One can alternatively consider other error structures, for instance, with $r_e(\Sigma_W) = O(1)$, in which case the above reasoning leads to the choice $\mu_n \gtrsim np$. However, this would limit the range of K, up to $\bar{K} = O(1)$ in (2.21), while our interest is in factor regression models with dimensions allowed to grow with n.

Proposition 20 in conjunction with Corollary 18 immediately leads to the following risk bound of PCR-$\widetilde{s}$. It coincides with the bound for PCR-$\widehat{s}$ in display (2.20) of Corollary 19.

Corollary 21. *Let $\widetilde{s}$ be defined in (2.21) with $\mu_n = c_0(n+p)$ for some absolute constant $c_0 > 0$. For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \le Cn/\log n$ such that (X, Y) follows sG-FRM(θ), $\log p \le cn$, $K \le \bar{K}$ and (2.22) holds, for some positive constants $c = c(\gamma_w)$ and $c' = c'(\gamma_w)$, we have*

$$\mathbb{P}_\theta \left\{ \mathbb{R}(\mathbf{U}_{\widetilde{s}}) - \sigma^2 \lesssim (K + \log n)\frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1} \beta \right\} \ge 1 - O(n^{-1}). \tag{2.27}$$

2.3.2 Existing Results on PCR

Due to the popularity and simplicity of PCR, its prediction properties under the factor regression model have been studied for nearly two decades. Most existing theoretical results, discussed below, are asymptotic in n and p and, to the best of our knowledge, have been established for a model of known dimension K, or when K is identifiable under additional restrictions on the parameter space, and can be consistently estimated.

The fact that PCR prediction, under the factor regression model with known or identifiable K, has asymptotically vanishing excess risk only when both p

and n grow to ∞ is a well known result. This can already be seen from our derivation (2.10) above, which shows that a necessary condition for prediction with vanishing excess risk, under factor regression models with well conditioned Σ_W, is $\|\Sigma_W\|_{\mathrm{op}}\beta^\top(A^\top A)^{-1}\beta \to 0$, which can be met when $p \to \infty$, as explained below.

This phenomenon was first quantified in [86], where it is shown that

$$\widehat{Y}^*_{\mathbf{U}_K} - Z_*^\top\beta = o_p(1) \ \text{ as } \ n, p \to \infty.$$

This result is the most closely related to ours, and we discuss it in detail below. We also mention that several later works, for instance [8] and [42], provided explicit convergence rates and inferential theory for the *in-sample* prediction error $\widehat{\boldsymbol{Y}} - \boldsymbol{Z}\beta$, whereas in this work we study out-of-sample performance. For completeness, we comment on these related, but not directly comparable, results in Appendix B.5.

In addition to being asymptotic in nature, the results in [86], and also those regarding the in-sample prediction accuracy, are established under the following set of conditions: $K = O(1)$, $\|\beta\|^2 = O(1)$, $\|\Sigma_W\|_{\mathrm{op}} = O(1)$, as $p \to \infty$, and

$$\frac{1}{p}A^\top A \to \boldsymbol{I}_K, \text{ as } p \to \infty, \quad \Sigma_Z \text{ is a diagonal matrix with distinct diagonal entries.} \tag{2.28}$$

These conditions serve as identifiability conditions for $\theta = (K, \beta, A, \Sigma_Z, \Sigma_W, \sigma^2)$ [86]. Condition (2.28) further implies that, for some constants $0 < c \le C < \infty$,

$$p \lesssim \lambda_K(AA^\top) \le \lambda_1(AA^\top) \lesssim p, \quad c \le \lambda_K(\Sigma_Z) \le \lambda_1(\Sigma_Z) \le C. \tag{2.29}$$

In contrast, our Corollaries 18, 19 and 21 are non-asymptotic statements, which hold for any finite K, n and p, where K is allowed to depend on n, with $K \log n \lesssim n$. Consequently, $\|\beta\|_2^2$ and $\lambda_1(\Sigma_Z)$ are also allowed to grow with n.

Furthermore, our conditions on the signal $\lambda_K(A\Sigma_Z A^\top)$ are much weaker than (2.29) to derive the risk bound of PCR-K. To see this, and for a transparent comparison, suppose $\|\Sigma_W\|_{\text{op}} \lesssim 1$ and $\lambda_K(\Sigma_Z) \geq c$. Then from Remark 6 we only require a condition much weaker than $\lambda_K(AA^\top) \gtrsim p$ of [86] given in (2.29) above, namely

$$\lambda_K(AA^\top) \gtrsim 1 + \frac{p}{n}.$$

Finally, the results in [86] are established for the unique θ under additional restrictions of the parameter space discussed above, whereas our results are established for any θ with $K \log n \lesssim n$ such that (X, Y) satisfying sG-FRM(θ), without requiring θ to be identifiable. In particular, our results hold for any identifiable θ that further satisfies (2.28).

We conclude our comparison by giving the bound implied by our Corollary 19, should the more stringent conditions (2.29) be met. Since (2.29) implies that $\widehat{s} = K$ with high probability from Remark 6, Corollary 19 immediately yields, with probability $1 - O(n^{-1})$,

$$\mathbb{R}(\mathbf{U}_K) - \sigma^2 \lesssim \frac{\log n}{n}\sigma^2 + \frac{\|\Sigma_W\|_{\text{op}}}{p} + \frac{\|\Sigma_W\|_{\text{op}}}{n},$$

and thus, as in [86],

$$\mathbb{R}(\mathbf{U}_K) - \sigma^2 = o_p(1)$$

when $p, n \to \infty$ and $\|\Sigma_W\|_{\text{op}} = O(1)$.

2.4 Analysis of Alternative Prediction Methods

In this section we illustrate the usage of the main Theorem 17 to derive risk bounds under a factor regression model for two other prediction methods: Gen-

eralized Least Squares [33], as an example of another model agnostic predictor construction, and model-tailored prediction, in an instance of an identifiable factor regression model provided by the *Essential Regression* framework introduced in [22]. All proofs for this section are contained in Appendix B.2.3.

2.4.1 Prediction Risks of Minimum Norm Interpolating Predictors Under Factor Regression Models

In the recent paper [33], risk bounds were established under the factor regression model for the Generalized Least Squares (GLS) predictor, which corresponds to taking $\widehat{B} = \boldsymbol{I}_p$:

$$\widehat{Y}^*_{\boldsymbol{I}_p} = X_*^\top \boldsymbol{X}^+ \boldsymbol{Y}. \tag{2.30}$$

We recover as these results in Corollary 22 and Corollary 23 below, as further illustration of the application of our main theorem. Since $P_{\boldsymbol{I}_p} = \boldsymbol{I}_p$ and $P^\perp_{\boldsymbol{I}_p} = 0$, the application of Theorem 17 with $\widehat{\psi} = 0$ amounts to obtaining a lower bound on the smallest non-zero singular value of $\boldsymbol{X}$ to bound $\widehat{\eta}$.

We consider the low ($p < n$)- and high ($p > n$)-dimensional settings separately. In the former case, GLS reduces to the ordinary least squares (OLS) method. The following corollary states the prediction risk of the OLS under the factor regression model. The proof uses a standard random matrix theory result [**?**, see]Theorem 5.39]vershynin$_2$012*toshow*$\sigma_p^2(\boldsymbol{X}) \gtrsim \lambda_p(\Sigma_W)n$, which implies $\widehat{\eta} \gtrsim \lambda_p(\Sigma_W)$. Recall that $\kappa(\Sigma_W) := \lambda_1(\Sigma_W)/\lambda_p(\Sigma_W)$.

Corollary 22 (GLS: low-dimensional setting)**.** *Suppose $p \log n \le c_0 n$ for an absolute constant $c_0 \in (0,1)$. For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \le Cn/\log n$ and $\lambda_p(\Sigma_W) >$*

c such that $(X, Y) \sim$ *sG-FRM*(θ)*, one has*

$$\mathbb{P}_\theta \left\{ \mathbb{R}(\boldsymbol{I}_p) - \sigma^2 \lesssim \left(\frac{p + \log n}{n} \sigma^2 + \|\Sigma_W\|_{op} \beta^\top (A^\top A)^{-1} \beta \right) \kappa(\Sigma_W) \right\} \geq 1 - O(n^{-1}).$$

When p is much larger than n, the GLS becomes the minimum ℓ_2 norm interpolator [33], one method studied in the recent wave of literature on the generalization of overparameterized models with zero or near-zero training error [13–18, 33, 44, 51, 72, 77–79]. Theorem 17 can also be applied to recover a slightly modified form of the prediction risk bound from [33] in this case, which we state in the following corollary. Recall that $r_e(\Sigma_W) = \mathrm{tr}(\Sigma_W)/\|\Sigma_W\|_{\mathrm{op}}$ is the effective rank of Σ_W.

Corollary 23 (GLS: high-dimensional setting. Interpolating predictors.)**.** *For any* $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ *with* $K \leq Cn/\log n$ *such that* $(X, Y) \sim$ *sG-FRM*(θ)*, suppose* $\widetilde{W}$ *defined in Definition 2.2.1 has independent entries and* $r_e(\Sigma_W) > C'n$ *for some sufficiently large constant* $C' > 0$*. Then there exists* $c > 0$ *such that*

$$\mathbb{P}_\theta \left\{ \mathbb{R}(\boldsymbol{I}_p) - \sigma^2 \lesssim \frac{K + \log n}{n} \sigma^2 + \frac{n}{r_e(\Sigma_W)} \sigma^2 + \frac{r_e(\Sigma_W)}{n} \|\Sigma_W\|_{\mathrm{op}} \beta^\top (A^\top A)^{-1} \beta \right\} \geq 1 - c/n.$$

By Proposition 6 of [33], we have $\sigma_n^2(\boldsymbol{X}) \gtrsim \mathrm{tr}(\Sigma_W)$ with high probability when $r_e(\Sigma_W) \gtrsim n$. Corollary 23 thus follows from Theorem 17 with $\widehat{\psi} = 0$ and $\widehat{\eta} \gtrsim \mathrm{tr}(\Sigma_W)/n$ in the high-dimensional setting. A simplified version of the risk bound in Corollary 23, together with a comparison with PCR-k prediction, is presented in Section 2.4.3.

2.4.2 Prediction Under Essential Regression

Both Principal Component Regression and Generalized Least Squares are model-agnostic methods, in that they do not use explicit estimates of the model parameters $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ to perform prediction. In contrast, further assumptions can be placed on the factor model to make θ identifiable, in which case a direct estimate of A can be meaningfully constructed and used for prediction. The Essential Regression (ER) framework introduced in [22] provides an approach to do this.

Essential Regression is a particular factor regression model under which the latent factor Z becomes interpretable under additional model assumptions. Specifically, under model (2.1), one further assumes the following model specifications.

Assumption 4.

(A0) $\|A_{j\bullet}\|_1 \leq 1$ *for all* $j \in [p]$.

(A1) For every $k \in [K]$, *there exists at least two* $j \neq \ell \in [p]$, *such that* $|A_{j\bullet}| = |A_{\ell\bullet}| = e_k$.

(A2) There exists a constant $\nu > 0$ *such that*

$$\min_{1 \leq a < b \leq K} \left([\Sigma_Z]_{aa} \wedge [\Sigma_Z]_{bb} - |[\Sigma_Z]_{ab}| \right) > \nu.$$

(A3) The covariance Σ_W *of* W *is diagonal with bounded diagonal entries.*

The indices $i \in [p]$ satisfying $A_{i\bullet} = e_k$ are called *pure variables* and collected in the set I. We use $J = [p] \setminus I$ to denote all the variables that are *non*-pure.

Within the Essential Regression framework, the matrix A becomes identifiable

up to a signed permutation [26]. In fact, $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ can be further shown to be identifiable [22].

We explain how to construct predictors of Y tailored to a factor model, and elaborate on the predictor tailored to Essential Regression. Under any factor model (2.1), the best predictor of Y from Z is $Z^\top\beta$. However, since Z is not observable, this expression does not lend itself to sample level prediction. A practically usable expression for a predictor under the factor regression model can be obtained by the following reasoning. Using the Moore-Penrose inverse $A^+ := (A^\top A)^{-1}A^\top$ of the matrix A, we observe that model (2.1) implies

$$\bar{X} := A^+X = Z + A^+W.$$

The best linear predictor (BLP) of Z from $\bar{X}$ is given by

$$\widetilde{Z} = \mathrm{Cov}(Z, \bar{X})[\mathrm{Cov}(\bar{X})]^{-1}\bar{X} = \Sigma_Z\left(\Sigma_Z + A^+\Sigma_W A^{+\top}\right)^{-1}A^+X. \tag{2.31}$$

The simple observation that

$$\arg\min_\alpha \mathbb{E}[(Y - Z^\top\alpha)^2] = \beta = \arg\min_\alpha \mathbb{E}[(Y - \widetilde{Z}^\top\alpha)^2]$$

justifies predicting Y by $\widetilde{Y} = \widetilde{Z}^\top\beta$. Inserting the identity $\beta = \Sigma_Z^{-1}A^+\mathrm{Cov}(X, Y)$ simplifies $\widetilde{Y}$ to

$$\begin{aligned}\widetilde{Y}_A &= X^\top A^{+\top}\left(\Sigma_Z + A^+\Sigma_W A^{+\top}\right)^{-1}\Sigma_Z\beta \\ &= X^\top A\left[\mathrm{Cov}(A^\top X)\right]^{-1}\mathrm{Cov}(A^\top X, Y),\end{aligned}$$

motivating prediction based on a new data point X_* by

$$Y^*_{\widehat{A}} = X_*^\top\widehat{A}\left(\widehat{A}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{A}\right)^+\widehat{A}^\top\boldsymbol{X}^\top\boldsymbol{Y},$$

which has the general form (2.3) with $\widehat{B} = \widehat{A}$, with $\widehat{A}$ being an estimator of A tailored to the ER model, developed in [26]. We summarize the construction of $\widehat{A}$

in Appendix B.4 for completeness.

To analyze the prediction risk of $Y^*_{\widehat{A}}$ we will also need the following assumption on the covariance matrix Σ_Z, which plays the same role as the Gram matrix in classical linear regression with random design.

Assumption 5. *Assume $c \le \lambda_K(\Sigma_Z) \le \lambda_1(\Sigma_Z) \le C$ for some constants c and C bounded away from 0 and ∞.*

The prediction risk of $\widehat{Y}^*_{\widehat{A}}$ can be obtained via an application of Theorem 17, with the choice $\widehat{B} = \widehat{A}$. Since A is identifiable under the Essential Regression framework, the estimator $\widehat{A}$ can be compared directly with A and, as shown in [26],

$$\|\widehat{A} - A\|^2_{\text{op}} \le \|A_J\|_0 \log(n \vee p)/n \tag{2.32}$$

with high probability. The rows of the $p \times |J|$ submatrix A_J of A correspond to all the index set J of non-pure variables. The estimation bound (2.32) can be leveraged to obtain a small improvement in the risk bound by slightly adjusting the proof of Theorem 17. Using this approach, we obtain the following result by establishing, with high probability, that

$$\begin{aligned}
&\widehat{r} = K, \\
&\widehat{\eta} \gtrsim \lambda_K(A\Sigma_Z A^\top), \\
&\widehat{\psi} \lesssim \|A_J\|_0 \frac{\log(p \vee n)}{n} + \|\Sigma_W\|_{\text{op}} := \psi_n(A_J).
\end{aligned}$$

Theorem 24 (Prediction in Essential Regression)**.** *Suppose $(X, Y) \sim$ sG-FRM(θ) with $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ satisfying Assumptions 4 & 5, $K \le Cn/\log n$ and*

$$\lambda_K(A\Sigma_Z A^\top) \ge c \cdot \psi_n(A_J)$$

for some sufficiently small constant $c > 0$. Then, with probability at least $1 - O(n^{-1})$,

$$\mathbb{R}(\widehat{A}) - \sigma^2 \lesssim \frac{K + \log n}{n}\sigma^2 + \psi_n(A_J)\beta^\top (A^\top A)^{-1}\beta. \tag{2.33}$$

Remark 7.

1. We note that the bound (2.33) depends on $\|A_J\|_0$, which in turn depends on the *number* of non-pure variables, and the *sparsity* of the rows of A corresponding to these non-pure variables. The rate indicates that prediction based on $\widehat{A}$ will perform best when the number of pure variables is large, and any non-pure variable X_i, the ith component of X, only depends on a small number of latent variables. We give, in the following section, a simplified form of this bound, and compare this prediction scheme with the other methods discussed in this work.

2. The identifiable factor model $X = AZ + W$, with A satisfying Assumption 4, has been used in [26] to construct overlapping clusters of the components on X. The latent factors can be viewed as random cluster centers, while a sparse matrix A gives the cluster membership. From this perspective, and in light of the discussion leading up to the predictor construction, one can view $\mathbb{R}(\widehat{A})$ as the risk of predicting Y from predicted cluster centers, on the basis of data that exhibits a latent cluster structure with overlap.

2.4.3 Comparison of Simplified Prediction Risks

In this section we offer a comparison of the prediction risk of the predictors analyzed above. For a transparent comparison, we compare them under an identifiable factor regression model. To this end, we consider the Essential Regression

framework as a data generating mechanism under which we compare PCR-k, with known $k = K$, the GLS predictor ($\widehat{B} = \boldsymbol{I}_p$), and the Essential Regression predictor ($\widehat{B} = \widehat{A}$), based on Corollary 19, Remark 6, Corollary 23 and Theorem 24, respectively. The notation $a_n \lessapprox b_n$ stands for $a_n = O(b_n)$ up to a multiplicative logarithmic factor in n or p.

For ease of comparison, we consider the simplified setting in which $\lambda_K(A^\top A) \gtrsim p/K$,[3] $\|\beta\|_2 \leq R_\beta$ and $r_e(\Sigma_W) \asymp p$, and focus on the high-dimensional regime where $p > Cn$ for a large enough constant $C > 0$. We have

$$\begin{aligned}
\mathbb{R}(\mathbf{U}_K) - \sigma^2 &\lessapprox \frac{K}{n}\sigma^2 + \frac{K}{p}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 + \frac{K}{n}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 \\
\mathbb{R}(\widehat{A}) - \sigma^2 &\lessapprox \frac{K}{n}\sigma^2 + \frac{K}{p}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 + \frac{K\|A_J\|_0}{np}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 \\
\mathbb{R}(\boldsymbol{I}_p) - \sigma^2 &\lessapprox \frac{K}{n}\sigma^2 + \frac{n}{p}\sigma^2 + \frac{K}{n}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2
\end{aligned} \tag{2.34}$$

Since the Essential Regression predictor is an instance of model based prediction, we comment on when the two model agnostic predictors are competitive, under this particular model specification.

We begin with a comparison between $\mathbb{R}(\mathbf{U}_K)$ and $\mathbb{R}(\widehat{A})$, and note that the difference in their respective errors bounds depends on the sparsity of A_J. The risk bound on $\mathbb{R}(\mathbf{U}_K)$ is valid for any θ such that $(X, Y) \sim$ sG-FRM(θ), and is in particular valid for θ satisfying the additional Essential Regression constraints. Our results show that while PCR-K prediction is certainly a valid choice under this particular model set-up, it could be outperformed by the model tailored predictor. If each row of A_J is sparse such that $\|A_J\|_0 \asymp |J|$, then $\mathbb{R}(\widehat{A})$ has a faster rate. This advantage becomes considerable if $|J| = o(p)$, that is, in the presence

[3] This is met for instance when all X's are pure variables and the numbers of pure variables for all groups are balanced in the sense that $|I_k| \asymp |I|/K$. Another instance such that $\lambda_K(A^\top A) \gtrsim p/K$ holds with high probability is that $|I_k| \asymp |I|/K$ and the rows of A_J are i.i.d. realizations of a sub-Gaussian random vector whose second moment has operator norm bounded by $1/K$. The factor $1/K$ takes (A0) in Assumption 4 into account.

of a growing number of pure variables. However, if A_J is not sparse such that $\|A_J\|_0 \asymp |J|K$, and $|J| \asymp p$, then $\mathbb{R}(\widehat{A})$ has a slower rate than $\mathbb{R}(\mathbf{U}_K)$. Nevertheless, from a practical perspective, conditions on the sparsity of A ($\|A_J\|_0 \asymp |J|$) simply mean that not all p variables in the vector X contribute to explaining a particular Z_k, for each k, which is the main premise of Essential Regression. Furthermore, in this risk bound comparison, $\mathbb{R}(\widehat{A})$ corresponds to $\widehat{A} \in \mathbb{R}^{p \times \widehat{K}}$, for an appropriate, fully data dependent, estimator $\widehat{K}$ of the identifiable dimension K. In order to employ a fully data driven PCR prediction, corresponding to an estimated K, we would also need the delicate step of estimating it described in Section 2.3 above. The risk bound above will then hold under conditions discussed in Remark 6.

Finally, the much simpler GLS interpolating predictor has a bound that compares favorably to the other agnostic predictor, PCR-K, only when n/p is small enough, for instance, $p > n^2/K$. This extra term $\sigma^2 n/p$ in the bound for $\mathbb{R}(\boldsymbol{I}_p)$ compared to the bound for PCR-K, is due to the additional variance induced by the usage the full data matrix $\boldsymbol{X}$, as opposed to the first K principal components, which may already capture the majority of the signal.

2.5 Predictor Selection via Data Splitting

Whenever a factor regression model can be assumed to generate a given data set, but it is unclear what further model specifications are in place, one can, in principle, construct several predictors, some model agnostic and some tailored to prior beliefs. In this section we address the problem of choosing among a set of candidate predictors for a given data set that is assumed to be generated by a factor regression model. Suppose we have M linear predictors with respective

coefficients $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_M$ that we want to choose from. For ease of presentation, in this section assume n is divisible by 2. Let D_1 be a subset of $[n]$ with $|D_1| = n/2$, and let $D_2 = [n] \setminus D_1$. Define

$$\widehat{m} := \arg\min_{m \in [M]} \sum_{i \in D_2} (Y_i - X_i^\top \widehat{\alpha}_m)^2, \tag{2.35}$$

where for each $m \in [M]$, $\widehat{\alpha}_m$ is trained on the data set $\{(X_i, Y_i) : i \in D_1\}$ and is thus independent of $\{(X_i, Y_i) : i \in D_2\}$. We then use $\widehat{\alpha} := \widehat{\alpha}_{\widehat{m}}$ as our predictor, for which we establish the following oracle inequality, which is an adaptation of Theorem 2.1 from [92] to factor regression models and unbounded linear predictors. Moreover, we provide a high-probability statement, as opposed to a bound on the expected risk as in [92]. The proof is deferred to Appendix B.2.4.

Theorem 25. *Let $\widehat{\alpha} := \widehat{\alpha}_{\widehat{m}}$, where $\widehat{m}$ is defined in (2.35). Then for any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ such that $(X, Y) \sim$ sG-FRM(θ), there exist absolute constants $c, c' > 0$ and a constant $c_0 = c_0(\gamma_w, \gamma_z, \gamma_\varepsilon) > 0$ such that when $n > c \log(M)$ and for any $a > 0$,*

$$\mathbb{P}_\theta \Bigg\{ \mathbb{R}(\widehat{\alpha}) - \sigma^2 \le (1+a)^2 \min_{m \in [M]} \{\mathbb{R}(\widehat{\alpha}_m) - \sigma^2\} + C(a) \left(\sigma^2 \vee \max_{m \in [M]} \{\mathbb{R}(\widehat{\alpha}_m) - \sigma^2\} \right) \frac{\log(nM)}{n} \Bigg\} \ge 1 - c' n^{-1}, \tag{2.36}$$

where $C(a) = c_0(1+a)^3/a$.

In the bound above, the worst excess risk $\max_m \{\mathbb{R}(\widehat{\alpha}_m) - \sigma^2\}$ appears in the remainder term, which may appear unusual. Most model-selection oracle inequalities either are formulated as a bound on the empirical risk, or assume that the predictors are uniformly bounded, or both, and as a result do not contain a term of this form. The bound we give is for the prediction risk on new data, and

for unbounded loss and predictors, since $\sup_\alpha (X^\top \alpha - y)^2 = \infty$. For the bound to be useful, it thus must be the case that none of the M predictors has risk that grows too fast. In particular, if the risks of all M predictors are bounded above in high probability, then the second term in (2.36) will be $O(\log n/n)$ and thus typically subdominant.

As an illustration, we can use this data-splitting procedure with $M = 3$ and the three prediction methods discussed in Section 2.4.3. If the three excess risks in (2.34) are all $O(1)$, which is met under the conditions discussed in detail in Section 2.4.3, then the bound (2.36) becomes

$$\mathbb{R}(\widehat{\alpha}) - \sigma^2 \lesssim (1+a)^2 \min\left(\mathbb{R}(\mathbf{U}_K) - \sigma^2,\ \mathbb{R}(\widehat{A}) - \sigma^2,\ \mathbb{R}(\boldsymbol{I}_p) - \sigma^2\right) + C(a)\sigma^2 \frac{\log n}{n}.$$

We further confirm the ability of the data-splitting approach to adapt to the best-case risk via simulations in Section 2.6 below.

On a practical note, we remark that the splitting procedure can be repeated several times with random splits to obtain estimates $\widehat{\alpha}^{(1)}, \ldots, \widehat{\alpha}^{(N)}$ that can be used to construct the average $N^{-1} \sum_{i=1}^N \widehat{\alpha}^{(i)}$. This aggregate coefficient vector satisfies the same risk bound (2.36) by convexity of the loss, while this approach in practice could alleviate some of the bias induced by the choice of split for the data.

2.6 Simulations

In this section, we complement and support our theoretical findings with simulations, focusing on the prediction performance of candidate predictors under both the generic factor regression model and the Essential Regression framework.

Candidate predictors: We consider the following list of predictors:

[leftmargin = 8mm]PCR-$\widetilde{s}$ with $\widetilde{s}$ obtained from (2.21) with $\mu_n = 0.25(n + p)$; PCR-K: the principal component regression (PCR) predictor using the true K; PCR-ratio: PCR with k selected via the criterion proposed in [2, 68]; [4] GLS: the Generalized Least Squares predictor defined in (2.30); ER-A: the Essential Regression predictor with $\widehat{B} = \widehat{A}$ in (2.3);Lasso: implemented in glmnet with the tuning parameter chosen via cross-validation; Ridge: implemented in glmnet with the tuning parameter chosen via cross-validation; MS: the selected predictor from (2.35) in Section 2.5.

Both Lasso and Ridge are included for comparison. The Lasso is developed for predicting Y from X when we expect that the best predictor of Y is well approximated by a sparse linear combination of the components of X. Under our model specifications, the best linear predictor of Y from X is given by

$$X^\top \alpha^* = X^\top [\mathrm{Cov}(X)]^{-1} \mathrm{Cov}(X, Y) = X^\top \Sigma_W^{-1} A \left[\Sigma_Z^{-1} + A^\top \Sigma_W^{-1} A\right]^{-1} \beta,$$

where the last step follows from the factor model (2.1) and an application of the Woodbury matrix identity. Although α^* is not sparse in general, we observe that $\|\alpha^*\|_2^2 \leq \beta^\top [\Sigma_Z^{-1} + A^\top \Sigma_W^{-1} A]^{-1} \beta$. Hence its ℓ_2-norm may be small if $\|\Sigma_W\|_{\mathrm{op}} \beta^\top (A^\top A)^{-1} \beta$ is small. Our simulation design allows for these possibilities.

Data generating mechanism: We first describe how we generate Σ_Z, Σ_W, and β. To generate Σ_Z, we set $\mathrm{diag}(\Sigma_Z)$ to a K-length sequence from 2.5 to 3

[4] We have also implemented the selection criterion suggested by [9], but it had inferior performance, and is for this reason not included in our comparison here.

with equal increments. The off-diagonal elements of Σ_Z are then chosen as $[\Sigma_Z]_{ij} = (-1)^{(i+j)}([\Sigma_Z]_{ii} \wedge [\Sigma_Z]_{jj})(0.3)^{|i-j|}$ for all $i \neq j \in [K]$. Finally, Σ_W is chosen as a diagonal matrix with diagonal elements sampled from Unif$(1, 3)$, and β is generated with entries sampled from Unif$(0, 3)$.

Generating A depends on the modeling assumption. Under the factor regression model, we sample each entry of A independently from $N(0, 1/\sqrt{K})$. Under the Essential Regression setting, recall that A can be partitioned into A_I and A_J which satisfy Assumption 4. To generate A_I, we set $|I_k| = m$ for each $k \in [K]$ and choose $A_I = \boldsymbol{I}_K \otimes \mathbf{1}_m$, where $\otimes$ denotes the kronecker product. Each row $A_{j\bullet}$ of A_J is generated by first randomly selecting its support with cardinality s_j drawn from $\{2, 3 \ldots, \lfloor K/2 \rfloor\}$ and then by sampling its non-zero entries from Unif$(0, 1/s_j)$ with random signs. In the end, we rescale A_J such that the ℓ_1 norm of each row is no greater than 1.

Finally, we generate the $n \times K$ matrix $\boldsymbol{Z}$ and the $n \times p$ noise matrix $\boldsymbol{W}$ whose rows are i.i.d. from $N_K(0, \Sigma_Z)$ and $N_p(0, \Sigma_W)$, respectively. We then set $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$ and $\boldsymbol{Y} = \boldsymbol{Z}\beta + \boldsymbol{\varepsilon}$ where the n components of $\boldsymbol{\varepsilon}$ are i.i.d. $N(0, 1)$.

For each setting, we generating 100 repetitions of $(\boldsymbol{X}, \boldsymbol{Y})$ and record their corresponding results. The performance metric is based on the new data prediction risk. To calculate it, we independently generate a new data set $(\boldsymbol{X}_{new}, \boldsymbol{Y}_{new})$ containing n i.i.d. samples drawn according to our data generating mechanism. The prediction risk of the predictor $\widehat{\boldsymbol{Y}}_{new}$ is calculated as $\|\widehat{\boldsymbol{Y}}_{new} - \boldsymbol{Z}_{new}\beta\|^2/n$.

2.6.1 Prediction Under the Factor Regression Model

We compare the performance of PCR-$\widetilde{s}$, PCR-K, PCR-ratio, GLS, Lasso, Ridge and MS by varying p, K and the signal-to-noise ratio (SNR) ξ defined in (2.8), one at a time. The MS predictor is based on (2.35) over all the aforementioned methods.

We first set $n = 300$, $K = 5$ and vary p from $\{100, 300, 700, 1500, 3000, 5000\}$, then choose $n = 300$, $p = 500$ and vary K from $\{3, 5, 10, 15, 20\}$. The prediction risks of different predictors for these two settings are shown in Figure 2.1. Since both PCR-$\widetilde{s}$ and PCR-ratio consistently select the true K, we only present the result for PCR-K.

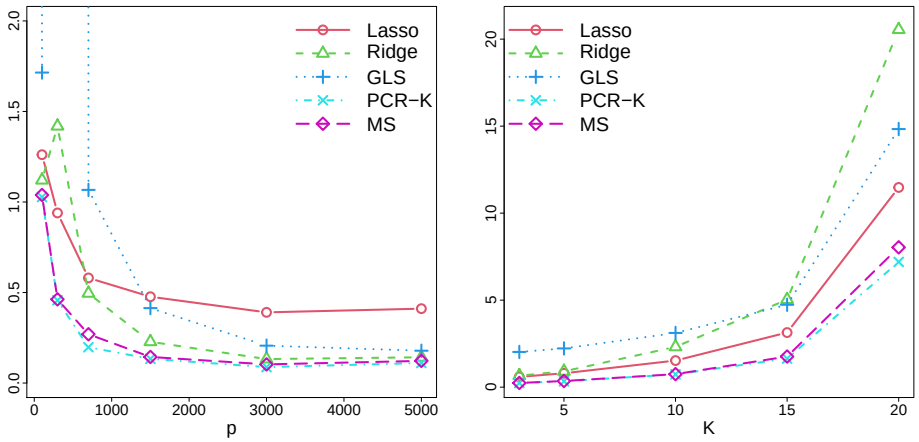

Figure 2.1: Prediction risks of different predictors under the factor regression model as p and K vary separately

Results: Overall, it is clear that the MS predictor selects the best predictor in almost all settings, corroborating Theorem 25. Meanwhile, PCR-K has the best performance in all settings as it is tailored to the factor regression model.

From the first panel, all methods perform better as p increases (with excep-

tions given to GLS and Ridge when $p \approx n = 300$). This contradicts the classical understanding that having more features increases the degrees of freedom of the model, hence inducing larger variance. By contrast, in our setting, increasing the number of features provides information that can be used to predict A. This can be seen from the minimal excess risk in Lemma 16 by noting that $\lambda_K(A^\top A)$ increases as p increases. This phenomenon has been observed in the classical factor (regression) model, see, for instance, [8, 10, 11, 42, 86] and the references therein.

Perhaps more interestingly, when p is much larger than n, GLS and Ridge have performance similar to PCR-K. This demonstrates our conclusions in Section 2.4.3 that GLS and PCR-K are comparable when $p \gg n$. We also note from our simulation that Ridge tends to select near-zero regularization parameter when $p \gg n$, whence Ridge essentially reduces to GLS [51]. In contrast to GLS and Ridge, the performance of Lasso stops improving after $p > 2500$. When p is moderately large (say $p < 1000$), GLS and Ridge have larger errors than PCR-K and Lasso. In particular, if p is close to n, the error of GLS diverges, a phenomenon observed in [51], for example, under the linear model.

From the second panel, the prediction error for all methods deteriorates as K increases. This indicates that prediction becomes more difficult for large K, supporting our results in Sections 2.3 and 2.4. We also note that the performance of Ridge deteriorates faster than the other methods when K grows.

To further demonstrate how different predictors behave as the signal-to-noise ratio (SNR) changes, we multiply A by a scalar α chosen within $\{0.1, 0.13, 0.16, \cdots, 0.37, 0.40\}$. We set $n = 300$, $p = 500$ and $K = 5$. For each

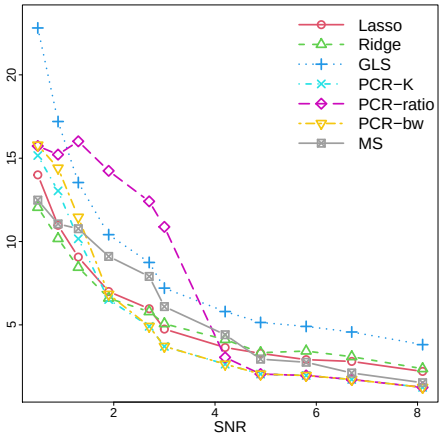

Figure 2.2: Prediction risks of different predictors under the factor regression model as SNR varies

α, we calculate the SNR and plot the prediction risks of each predictor in Figure 2.2.

Results: As expected, all methods perform worse as the SNR decreases. MS has consistently selected the (near) best predictor. When the SNR is small (less than 2), Ridge has the best performance. As soon as the SNR exceeds 2, PCR-K and PCR-$\widetilde{s}$ start to outperform the other methods. In terms of selecting K, when the SNR is larger than 2, PCR-$\widetilde{s}$ starts estimating K consistently whereas PCR-ratio fails until the SNR is greater than 4. Both PCR-$\widetilde{s}$ and PCR-ratio tend to underestimate K in the presence of a small SNR. However, PCR-$\widetilde{s}$ selects $\widetilde{s}$ closer to K than PCR-ratio, leading to better performance. Moreover, the loss due to using $\widetilde{s} < K$ by PCR-$\widetilde{s}$ is not significant, in line with Corollary 21 and Remark 6.

2.6.2 Prediction Under the Essential Regression Model

We compare all the predictors when data is generated from an Essential Regression model. To vary p and K individually, we first set $n = 300$, $K = 5$, $m = 5$

and choose p from $\{100, 300, 500, 700, 900\}$, then fix $n = 300$, $p = 500$, $m = 5$ and vary K in $\{3, 5, 10, 15, 20\}$. The prediction risks of different predictors are shown in Figure 2.3. PCR-$\widetilde{s}$ and PCR-ratio are not included as they have almost the same performance as PCR-K. As it was demonstrated under the factor regression setting that GLS is outperformed by the other predictors when p is not large enough, we also excluded its performance from the plot.

Summary: We observe the same phenomenon as before, that is: (1) all predictors benefit from large p; (2) as K increases, the performance of all predictors deteriorate. Furthermore, the model-based ER predictor has similar performance as the model-free PCR predictor when K is small. The advantage of ER over PCR enlarges as K grows. This is aligned with our theoretical findings in Section 2.4.3 that ER benefits from the sparsity of A_J, because our data generating mechanism ensures that the larger K is, the sparser A_J becomes.

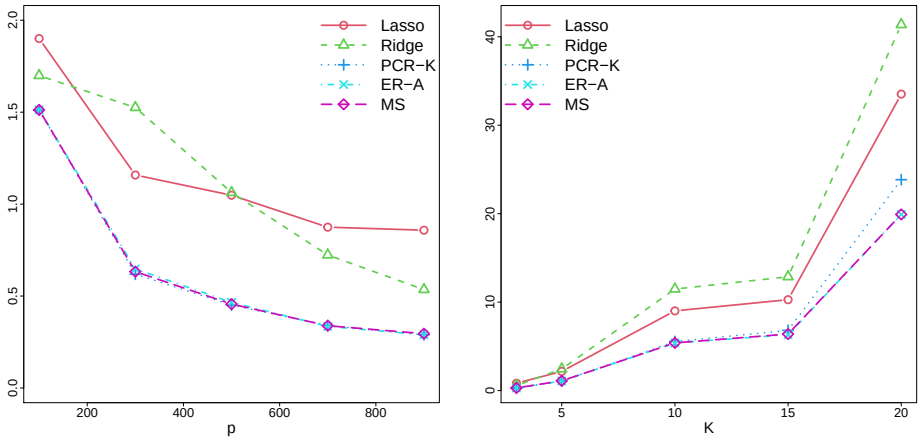

Figure 2.3: Prediction risks of different predictors under the Essential Regression model as p and K vary separately

• CHAPTER 3

A HIERARCHICAL DISTANCE BETWEEN CORPORA USING OPTIMAL TRANSPORT OF TOPIC-BASED CLUSTER DISTRIBUTIONS

3.1 Introduction

In this work we propose a new method for comparing pairs of corpora of documents generated from sets of latent topics, using a hierarchical Wasserstein-type metric, tailored to this problem. Although the Wasserstein distance, and its many modifications, are routinely used for *two-sample* statistical comparisons, e.g. [37,74,81,89], very little is known about the construction and analysis of its variants that can be readily applied to a *two-ensemble* comparison, where each ensemble is itself a collection of independent, but not necessarily identically distributed, samples and, moreover, can be sampled from high-dimensional distributions. We treat this problem here, in the topic model setting, and propose a distance for comparing ensembles of high-dimensional, discrete, distributions.

We assume that we observe two corpora of n and n' documents, respectively. We assume the corpora share a common dictionary of size p, which can be taken to be the set of unique words in the two corpora.

For the first corpus, each document $i \in [n] := \{1, \ldots, n\}$ is modelled as a sample of N_i words drawn from a discrete distribution $\Pi^{(i)}$ over the p words in the dictionary. We observe the p-dimensional word-count vector $Y^{(i)}$ for each document $i \in [n]$, where we assume

$$Y^{(i)} \sim \text{Multinomial}_p(N_i, \Pi^{(i)}).$$

The observed word frequencies, for all n samples, are collected in the $p \times n$

word-frequency matrix $\boldsymbol{X}$.

The topic model assumption is that the matrix of expected word frequencies in the corpus, $\mathbb{E}(\boldsymbol{X}) =: \mathbf{\Pi} \coloneqq (\Pi^{(1)}, \ldots, \Pi^{(n)})$ can be factorized as $\mathbf{\Pi} = A\boldsymbol{T}$. Here A represents the $p \times K$ matrix of conditional probabilities of a word, given a topic, and therefore each column of A belongs to the p-dimensional probability simplex

$$\Delta_p \coloneqq \{x \in \mathbb{R}^p \mid x \geq \mathbf{0},\ \mathbf{1}_p^\top x = 1\}.$$

The notation $x \geq \mathbf{0}$ specifies that $x_j \geq 0$ for each $j \in [p]$, and $\mathbf{1}_p$ is the vector of all ones. In particular, the kth column $A_{\cdot k}$ is a distribution over words in the dictionary conditional on topic k, with components

$$A_{ik} = \mathbb{P}(\text{word } i | \text{topic } k) \quad \forall i \in [p].$$

The $K \times n$ matrix $\boldsymbol{T} \coloneqq (T^{(1)}, \ldots, T^{(n)})$ collects the probability vectors $T^{(i)} \in \Delta_K$, the simplex in $\mathbb{R}^K$. The entries of $T^{(i)}$ are probabilities with which each of the K topics occurs within document i, for each $i \in [n]$, and are typically sparse.

In this setting, each generating distribution $\Pi^{(i)}$ is therefore a discrete mixture of $A_{\cdot k}$, with weights $T_k^{(i)}$, for $k \in [K]$ and $i \in [n]$.

For the second corpus, the same assumptions hold where $n, N_i, Y^{(i)}, \mathbf{X}, \mathbf{\Pi}, A, K, \boldsymbol{T}$ are replaced by $n', N_i', Y'^{(i)}, \mathbf{X}', \mathbf{\Pi}', A', K', \boldsymbol{T}'$. Note in particular that we allow for $K \neq K'$, so the corpora can have different numbers of topics, and $A \neq A'$, in which case the topics themselves (as represented by a column of A or A') can differ between the corpora. Furthermore, the topic proportions per document, collected in $\boldsymbol{T}$ and $\boldsymbol{T}'$, can be different, even if the corpora cover the same topics. The number of topics K and K' are not known prior to estimation, and are allowed to depend on and grow with the size of the corpus.

We will employ ideas drawn from optimal transport to define a distance between corpora, noting the added challenge posed by having independent, but not identically distributed high-dimensional samples. In the latter case, at the conceptual level, one could employ appropriate estimators of versions of the Wasserstein distance, proposed to alleviate the curse of dimensionality in, for instance, [45], but they employ empirical estimates of the common distributions underlying each ensemble. Since in a topic model setting each sample in the ensemble has its own distribution, different techniques must be developed. We propose the following strategy, stated at the population-level below, and expanded upon in Section 3.2.

(1) Reduction step: We cluster each corpus by topic. Document i of corpus 1 belongs to cluster a if $T_a^{(i)} > 0$, for each $a \in [K]$, thereby reducing this corpus to a set of K clusters. Repeating the procedure, the second corpus is reduced to a collection of K' clusters. The resulting clusters, of each corpus, will typically be overlapping, as documents can cover in detail, or only touch upon, multiple topics.

(2) Representation step. We represent each corpus as a discrete distribution over cluster centers. To this end, in Corpus 1, to each cluster $a \in [K]$ we associate, in (3.6), an appropriately defined weight $\theta_a > 0$, with $\sum_{a=1}^{K} \theta_a = 1$, and a center $\boldsymbol{\mathcal{T}}^{(a)}$ (Section 3.2.1). Each cluster center itself is viewed as a discrete measure, supported on K points in Δ_p consisting in the K discrete mixture components $A_{\cdot k} \in \Delta_p$:

$$\boldsymbol{\mathcal{T}}^{(a)} = \sum_{k=1}^{K} \boldsymbol{\mathcal{T}}_k^{(a)} \delta_{A_{\cdot k}}, \tag{3.1}$$

where δ_u is the Dirac measure concentrated on u, and $\boldsymbol{\mathcal{T}}_k^{(a)}$ is defined in (3.7). The set of all discrete measures on at most K points in Δ_p is denoted by $\mathcal{D}_{K,p}$. We use

the same reasoning to define $\mathcal{T}'^{(a)}$ and $\mathcal{D}_{K',p}$, for Corpus 2.

We then represent Corpus 1 as a discrete measure $\boldsymbol{\theta}$ supported on the K points $\mathcal{T}^{(a)} \in \mathcal{D}_{K,p}$:

$$\boldsymbol{\theta} := \sum_{a=1}^{K} \theta_a \delta_{\mathcal{T}^{(a)}}. \tag{3.2}$$

The set of all discrete measures on at most K points in $\mathcal{D}_{K,p}$ is denoted by $\mathcal{D}_{K,K}$, and so $\boldsymbol{\theta} \in \mathcal{D}_{K,K}$. Similarly, Corpus 2 is represented by $\boldsymbol{\theta}' =: \sum_{a=1}^{K'} \theta'_a \delta_{\mathcal{T}'^{(a)}}$, and $\boldsymbol{\theta}' \in \mathcal{D}_{K',K'}$. Note that both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ lie in $\mathcal{D}_{K\vee K',K\vee K'}$.

(3) A distance between corpora. With these ingredients in place, we define the distance between a pair of corpora as a distance between their probabilistic representations $D^{\text{corpora}} : \mathcal{D}_{K\vee K',K\vee K'} \times \mathcal{D}_{K\vee K',K\vee K'} \to \mathbb{R}_+$ given by

$$D^{\text{corpora}}(\boldsymbol{\theta}, \boldsymbol{\theta}') := W_1(\boldsymbol{\theta}, \boldsymbol{\theta}'; d^{\text{cluster}}), \tag{3.3}$$

where W_1 refers to the 1-Wasserstein distance between the two distributions, and is defined in Section (3.2.2). To complete the definition of the distance we need to specify a distance between $\mathcal{T}^{(a)}$ and $\mathcal{T}'^{(b)}$, for each $a \in [K]$, $b \in [K']$, denoted by

$$d^{\text{cluster}} : \mathcal{D}_{K\vee K',p} \times \mathcal{D}_{K\vee K',p} \to \mathbb{R}_+.$$

We let $d^{\text{cluster}}(a, b) := d^{\text{cluster}}(\mathcal{T}^{(a)}, \mathcal{T}'^{(b)})$ and, owing to the probabilistic representation (3.1), we once again use the Wasserstein distance to define

$$d^{\text{cluster}}(a, b) := W_1(\mathcal{T}^{(a)}, \mathcal{T}'^{(b)}; d^{\text{mix}}), \tag{3.4}$$

where $d^{\text{mix}}(k, l)$ is a distance between discrete mixture pairs $(A_{\cdot k}, A'_{\cdot l}) \in \Delta_p \times \Delta_p$, for $(k, l) \in [K] \times [K']$. While any distance between discrete probability vectors can be employed at this step, for computational simplicity and to widen the applicability of the final corpora distance, we work with the total variation distance and, with

slight abuse of notation, we define

$$d^{\text{mix}}(k, l) := \frac{1}{2}\|A_{\cdot k} - A'_{\cdot l}\|_1. \tag{3.5}$$

Displays (3.3) - (3.5) show the hierarchical construction of our proposed distance and we henceforth refer to (3.3) as the *hierarchical Wasserstein corpus distance* (HWCD).

3.1.1 Existing Results and Our Contribution

Other hierarchical versions of the Wasserstein distance are scattered throughout the literature, and used for diverse purposes. For instance, they can be important technical tools in theoretical Bayesian analyses [80], or used to define multi-level clustering schemes [52], by extending what has become a classical usage of the Wasserstein distance in the context of K-means clustering [84].

In the context of topic models, hierarchical variants of the Wasserstein distance have been used recently with empirical success in [94], and also received sharp theoretical treatment in [21], but for the problem of *document* comparison, in *one* corpus, which is an instance of a two-sample comparison, unlike the two-ensemble comparison treated here.

Earlier applications of other versions of hierarchical optimal transport that are closer to the problem tackled in this work can be found in [36], which provides a successful empirical comparison of two general, given, non-overlapping clustering schemes, but does not provide accompanying supporting theory.

The problem of corpora comparison, with theoretical guarantees, has not been studied, to the best of our knowledge, but a very limited number of empirical

studies, based on approaches that differ from our proposal, exist, mainly in the linguistics literature. Quantitative measures of corpora similarity were first proposed in [65]. A χ^2 measure based on word frequencies was found by [65] to perform the best at recovering ground-truth similarity in a set of so-called "Known Similarity Corpora".

The only topic-based similarity measure in the literature that we are aware of that is specifically dedicated to corpora comparison is [46]. This empirical distance is not based on estimating a population-level target and is without supporting theory. It also does not account for the relative distance between topics. In contrast, our distance is defined at the population level as a metric on a space of probabilistic representations of corpora. This allows for theoretical guarantees on its estimation, which we provide. Furthermore, by using the Wasserstein distance, our method incorporates the relative distance between topics. Due to our optimal transport approach, our method also provides a transport plan between the topics in a pair of corpora, aiding in an interpretable comparison of the corpora (See Figure 3.2 for a real-data example).

In light of existing results, our contribution is summarized below.

(1) We provide a new, principled, construction of a distance between two ensembles of independent, but not identically distributed, discrete samples, when each ensemble follows a topic model. Our proposal is a hierarchical Wasserstein distance, that can be used for document corpora comparison, or any other data sets following topic models. All the details are given in Section 3.2.

(2) We provide computationally tractable estimates of the distance, as well as accompanying finite sample error bounds, in Theorem 26. The final rate

cumulates the minimax-optimal error rate $\sqrt{K/n}$ incurred by estimating the mixture weight vectors $T^{(i)}$, $i \in [n]$, with the minimax-optimal error rate $\sqrt{\frac{\|A\|_0}{nN}}$ of estimating the word-topic matrix A, stated here for simplicity with $N_i = N$, for all $i \in [n]$. The second corpus contributes similar error bounds to the final corpus-distance estimate rate. The norm $\|A\|_0$ is the ℓ_0 norm of the matrix A, counting the number of elements in its support, and equals pK for a fully dense matrix, but can be much smaller when A is sparse. The latter is expected to hold in realistic scenarios since, given a topic, many words in a large dictionary will have a very small, or zero, conditional probability of occurrence. Consequently, our distance error bound shows that the distance between two corpora can be estimated accurately even if they consist in short documents (small N) as long as the size of each corpus n, n', is relatively large, and that the latter itself can be relaxed as long as A is sparse. The details are given in Section 3.3. An application to the comparison of news sources is provided in Section 3.4.

3.2 A hierarchical Wasserstein distance between corpora

In this section we follow the program laid out in the introduction and expand on each of the steps that lead up to the construction of our distance. We begin by detailing the construction of the probabilistic representation of each corpus.

3.2.1 A representation of a document corpus as a discrete distribution on cluster center distributions

In this section we use the notation associated with Corpus 1; that for Corpus 2 will follow by analogy. We recall that, for each $i \in [n]$, the components of the vector $T^{(i)} \in \Delta_K$ are the proportions in which each of the K topics in the corpus is covered by document i, and thus we expect $T^{(i)}$ to be sparse.

We begin by grouping the n documents by the topics they respectively address, including in a group a all documents i for which $T_a^{(i)} > 0$, for each $a \in [K]$, thereby creating K possibly overlapping clusters. To each topic-based cluster $a \in [K]$ we associate two objects: (1) A mass $\theta_a \in [0, 1]$, designed to reflect the relative weight topic a has in the entire corpus; and (2) A cluster center $\boldsymbol{\mathcal{T}}^{(a)}$, a discrete measure supported on K points in Δ_p detailed below.

We assign mass to a topic-based cluster $a \in [K]$ in a natural way:

$$\theta_a := \frac{1}{n} \sum_{i=1}^{n} T_a^{(i)}, \tag{3.6}$$

the average proportion in which topic a is represented in the entire corpus. It may appear that we have been too liberal in our cluster construction, in that we allowed a topic-based cluster a to include documents that may barely touch upon a topic, such that $T_a^{(i)} > 0$, but possibly very close to zero. However, we adjust for this in the definition of the cluster centers $\boldsymbol{\mathcal{T}}^{(a)}$, for each $a \in [K]$.

Recognizing that reducing a topic-based cluster to one representative could, in principle, lose some of the within-cluster variability, we do not only represent the cluster center by a vector of numerical values, but also add to them respective amounts of mass, to encode their potential variation. We are therefore led

naturally to viewing a cluster center as being itself a discrete measure.

In a topic model framework, the variation among documents is induced by a variation in their topic distributions, $T^{(i)} \in \Delta_K$, the entries of which place mass on each of the K topics. Each topic gives rise to word frequencies specific to that topic, collected in the word-topic vectors $A_{\cdot k}$. This allows for the identification of a topic k with $A_{\cdot k} \in \Delta_p$, a process that has been shown to lead to empirical success in [94], and vetted theoretically in [21]. We follow the same line of reasoning here, and will represent a cluster center as a discrete measure supported on K points, $A_{\cdot k} \in \Delta_p$, $k \in [K]$. We then define the mass it places on a point k as a weighted average of document-specific topic proportions $T_k^{(i)}$, over the entire corpus,

$$\boldsymbol{\mathcal{T}}_k^{(a)} := \sum_{i=1}^{n} \gamma_i^{(a)} T_k^{(i)}. \tag{3.7}$$

The weight $\gamma_i^{(a)}$ is the proportion of topic a in document i, relative to the proportion of topic a in the full corpus, and is defined by

$$\gamma_i^{(a)} := \frac{T_a^{(i)}}{\sum_{j=1}^{n} T_a^{(j)}}, \qquad \forall i \in [n]. \tag{3.8}$$

In the limit case in which each document in the corpus pertains exactly to one topic, let $M_a = |\{j \in [n] : \ T_a^{(j)} > 0\}|$ be the number of documents on topic a. Then for each topic $a \in [K]$ and document $i \in [n]$, $\gamma_i^{(a)} = 1/M_a$ if document i contains topic a, and $\gamma_i^{(a)} = 0$ otherwise, and thus $\gamma^{(a)}$ is the uniform distribution over documents containing topic a. In general, the definition (3.8) takes into account the potential topical complexity of each document. With these ingredients, a cluster center is associated with the discrete measure $\boldsymbol{\mathcal{T}}^{(a)}$ given by (3.1) in the introduction.

To complete the representation process for Corpus 1, we use mass θ_a given by (3.6) and cluster center $\boldsymbol{\mathcal{T}}^{(a)}$ given by (3.1), in conjunction with (3.7) and (3.8),

to obtain the discrete measure $\boldsymbol{\theta}$ given by display (3.2) of the introduction. We repeat the process to associate the measure $\boldsymbol{\theta}'$ with Corpus 2.

3.2.2 A hierarchical Wasserstein distance for two-ensemble comparison

Once each corpus is represented as a discrete distribution, we define a distance between corpora as a distance between their probabilistic representations $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Hierarchically employing distances rooted in optimal transport allows us to take full advantage of the geometry of the underlying probability space of discrete measures $\mathcal{D}_{K\vee K',p}$ that contains the support points of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$.

To see how, we first recall the definition of the Wasserstein distance for generic discrete distributions r and s with finite support on a generic metric space $(\mathcal{X}, d)$. The 1-Wasserstein distance between r and s is defined by

$$W_1(r, s; d) := \inf_{w\in\Sigma_W(r,s)} \sum_{x\in\text{supp}(r),y\in\text{supp}(s)} w(x, y) \cdot d(x, y), \tag{3.9}$$

where $\Sigma_W(r, s)$ denotes the set of all joint distributions (couplings) between r and s, with marginals r and s, and $\text{supp}(r)$, $\text{supp}(s)$ denote the support of r and s, respectively. We note that the minimizer w^* of 3.9 is called a *transport plan*, and can be use for interpreting how the measures r and s align.

We specialize this to distributions $\boldsymbol{\theta}, \boldsymbol{\theta}'$ and $d = d^{\text{cluster}}$ to obtain our proposed D^{corpora} (3.3) and to distributions $\boldsymbol{\mathcal{T}}^{(a)}$, $\boldsymbol{\mathcal{T}}'^{(b)}$ and $d = d^{\text{mix}}$ (3.5) to calculate the distance d^{cluster} (3.4), all defined in the introduction. We note that the overall computing effort involved in calculating D^{corpora} reduces to two optimization

problems in dimension $K \times K'$, a much reduced dimension relative to the ambient dimensions of the problem.

3.2.3 The discriminating power of the distance between corpora with varying topical content and topic coverage

In this sub-section we offer a simulation study, at the population level, to aid with the intuitive understanding of the properties of the new distance D^{corpora}. We focus on illustrating the discriminating power of the corpora-distance as a function of the different aspects in which two corpora satisfying a topic model can differ: (I) In terms of their respective word-topic matrices A and A'; and (II) In terms of their topic distributions $(T^{(i)})_{i\in[n]}$ and $(T'^{(i)})_{i\in[n']}$. If two corpora differ in any combination of (I) and (II), their associated measures $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ will differ. We illustrate that the new D^{corpora} can capture these differences through a small numerical study.

In particular, we illustrate below that D^{corpora} has maximal value, 1, when the two corpora cover a disjoint set of topics, as expected of any bona fide corpora-distance and that, moreover, can distinguish between corpora on similar topics ($A \approx A'$), but which are covered in different proportions by the documents of each corpus, $(T^{(i)})_{i\in[n]} \neq (T'^{(i)})_{i\in[n']}$.

We generate A with $p = 5000$, $K = 10$ by sampling the entries of the top half of A iid from Unif$(0, 1)$, setting the entries in the lower half to zero, and normalizing so the columns of A sum to 1. Let B be a $p \times K$ matrix generated independently of A, with the upper half entries set to zero, lower half entries drawn iid from

Unif(0, 1), and columns normalized to sum to 1. Then, for $h \in [0, 1]$, let

$$A' = (1 - h) * A + h * B. \tag{3.10}$$

Define $\alpha_{:(K/2)} \in \Delta_K$ such that the first $K/2$ components equal $2/K$, and the remaining components are zero. Let $\alpha_{(K/2):} \in \Delta_K$ have the first $K/2$ components equal to zero, and the remaining components equal to $K/2$. We generate $T^{(1)}, \dots T^{(n)}$ iid such that

$$T^{(i)} \propto \alpha_{:(K/2)} + \sigma * \varepsilon^{(i)}, \tag{3.11}$$

where $\varepsilon^{(1)}, \dots, \varepsilon^{(n)} \sim \text{Dirichlet}(\mathbf{1}_K)$ is iid noise and $\sigma = 0.01$. Thus the first corpus is concentrated on the first $K/2$ topics. We generate $T'^{(1)}, \dots T'^{(n)}$ iid such that

$$T'^{(i)} \propto (1 - t) * \alpha_{:(K/2)} + t * \alpha_{(K/2):} + \sigma * \varepsilon'^{(i)}, \tag{3.12}$$

where again $\varepsilon'^{(1)}, \dots, \varepsilon'^{(n)} \sim \text{Dirichlet}(\mathbf{1}_K)$ is iid noise and $t \in [0, 1]$. When t is close to zero, the second corpus is also concentrated on the first $K/2$ topics, and the two corpora have similar topic coverage, in each document. As t increases, more weight is placed on the final $K/2$ topics in the second corpus, until, when t is close to 1, $T'^{(1)}, \dots T'^{(n')}$ are mostly supported on the final $K/2$ documents, making the corpora dissimilar in terms of their topic coverage.

In Figure 3.1, we plot the corpora distance (3.3) between the two corpora defined by $(A, (T^{(i)})_{i \in [n]})$ and $(A', (T'^{(i)})_{i \in [n']})$, respectively, as a function of the parameter h in (3.10), for four representative values of the parameter t in (3.12). When $h = 0$, on the far left of the plot, $A = A'$, and two corpora have identical topical content. When $t = 0$, they also have identical topic coverage, thus the distance is zero. As t increases, although the topics are the same, the second corpus places increasingly greater weight on the final $K/2$ topics, and the corpora distance increases, as seen in the figure. When $h = 1$, on the far right of the plot,

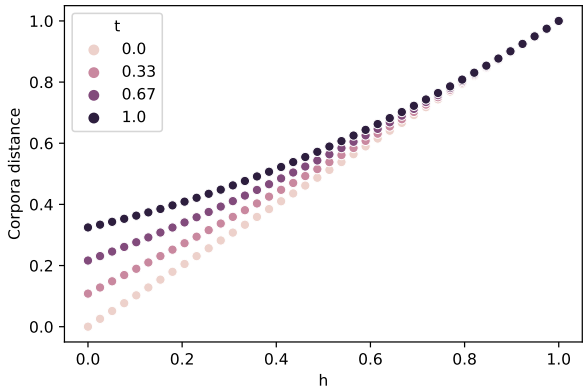

Figure 3.1: Corpora distance as a function of h in (3.10), for four representative values of the parameter t in (3.12).

$A' = B$; the topics in the two corpora then have disjoint support, resulting in a corpus distance of 1 regardless of topic coverage.

3.3 Estimation of the HWCD: methods and error bounds

We estimate D^{corpora} via plug-in estimates. The next sub-section gives our proposed estimates of the number of topics, K and K', of the word-topic matrices A and A', and of the topic distributions $T^{(i)}$ and $T'^{(i)}$, as well as their theoretical guarantees. For all estimated quantities we use the same notation we used for their population-level counterparts, to which we add the hat symbol. We then plug-in these estimates in (3.6) and (3.2) to obtain the estimator $\widehat{\boldsymbol{\theta}}$, and proceed similarly using (3.7), (3.8) and (3.1) to construct $\widehat{\boldsymbol{\mathcal{T}}}^{(a)}$, for each $a \in \widehat{K}$. We repeat the process for Corpus 2. Then, following (3.3), our final estimate is

$$\widehat{D}^{\text{corpora}}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}') := W_1(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}'; \widehat{d}^{\text{cluster}}), \tag{3.13}$$

where, using the same notation convention, and for $a \in [\widehat{K}], b \in [\widehat{K}']$,

$$\widehat{d}^{\text{cluster}}(a,b) := W_1(\widehat{\boldsymbol{\mathcal{T}}}^{(a)}, \widehat{\boldsymbol{\mathcal{T}}}'^{(b)}; \widehat{d}^{\text{mix}}), \tag{3.14}$$

with $\widehat{d}^{\text{mix}}$ defined as the Total Variation distance on Δ_p, similarly to d^{mix}. We use the hat to emphasize that it acts on support points $(\widehat{A}_{\cdot k})_{k\in[K']}$ and write $\widehat{d}^{\text{mix}}(a,b) := \frac{1}{2}\|\widehat{A}_{\cdot a} - \widehat{A}'_{\cdot b}\|_1$.

3.3.1 Error bounds on corpora-distance estimates

The error incurred in the estimation of D^{corpora} cumulates the errors induced by the estimation of the columns of the word-topic matrices (the mixture components) and of the topic-document distributions (their mixture weights). Minimax-rate optimal estimators of these quantities have only been developed very recently. We discuss them separately below and use them to obtain, in Theorem 26, our final distance error bound.

Estimators of word-topic matrices: The estimation of A (in Corpus 1, and similarly A' in Corpus 2) under topic models was originally studied within a Bayesian framework [31, 48], and variational-Bayes type approaches were further proposed to accelerate the computation of fully Bayesian approaches. We refer to [29] for an in-depth overview of this class of techniques. More recently, [3, 5, 7, 23, 24, 39, 62] studied provably fast algorithms for estimating A from a frequentist point of view. The common thread of these works, both theoretically and computationally, is the usage of what is known as the anchor word assumption, which assumes the existence of at least one word in the dictionary that is used in only one topic, and has been shown empirically to hold [38] in most large corpora for which topic models are reasonable modeling

tools. Furthermore, under this assumption, the topic model is identifiable, and the estimation of A is a well-posed problem. Theoretically validated estimation procedures that employ this identifiability assumption have been developed in [5, 28], when K is known, and extended to the case in which K is unknown, and also allowed to depend on the size n of the corpus, in [23]. The latter work provides consistent estimators of K, establishes the minimax lower bound for the estimation of A in $\|\ \|_{1,\infty}$ norm, and offers the first minimax-rate adaptive estimator in a regime in which K and p are allowed to depend on the sample sizes. In this paper we use an extension of this work, that further allows for the realistic scenario in which A is sparse.

We construct $\widehat{K}$ and $\widehat{A}$ using Algorithm 2 in [23] and Algorithm 1 in [24], which by their Theorem 2 and Corollary 3, is minimax-rate optimal. The error rate, that holds with high probability under the assumptions of these theorems, is

$$\min_{P \in \mathcal{H}} \|\widehat{A} - AP\|_{1,\infty} \lesssim \sqrt{\frac{\|A\|_0}{nN}}. \tag{3.15}$$

Here $\mathcal{H}$ is the set of $K \times K$ permutation matrices, and the symbol $\lesssim$ means that the inequality holds up to constants and possibly logarithmic factors. We also use this estimator for Corpus 2.

Estimators of the mixture weights: Computationally efficient methods, with theoretical guarantees, for estimating the topic distributions $T^{(i)}$, for $i \in [n]$, of one corpus, are scarce, with earlier results in [66], [6], restricted to known word-topic matrices A. Computationally efficient estimators of $T^{(i)}$ whose rate also reflects the error incurred by the estimation A have only been established very recently in [67] and [21], and include minimax-rate analyses. The latter work proposes a profile likelihood estimator, which we will also adopt here. The estimator

optimizes over $\Delta_{\widehat{K}}$ a multinomial likelihood function

$$\widehat{T}^{(i)} = \text{argmax}_{T \in \Delta_{\widehat{K}}} N \sum_{j=1}^{p} X_j^{(i)} \log\left(\widehat{A}_{j\cdot}^{\top} T\right), \tag{3.16}$$

for each i, and for given estimators $\widehat{A}$ and $\widehat{K}$. We will work with the estimator $\widehat{A}$ of [24], which attains the rate (3.15), and its associated $\widehat{K}$, which estimates K consistently. Then, Corollary 10 in [21] shows that

$$\min_{P \in \mathcal{H}} \|\widehat{T}^{(i)} - P^{\top} T^{(i)}\|_1 = O_{\mathbb{P}}\left(\sqrt{\frac{K \log p}{N}} + \sqrt{\frac{\|A\|_0 \log p}{nN}}\right), \tag{3.17}$$

for every i. The standard symbol $O_{\mathbb{P}}$ means that the rate in the right hand side holds with high probability.

We note that if A is known, the last term in the error bound is zero, and if an estimator with another rate is used, that rate will replace the second term above, which can be seen by inspecting the proof of Corollary 10 in [21]. Furthermore, this work also showed that, under appropriate conditions, $\widehat{T}^{(i)}$ can be exactly sparse, a remarkable property for an un-penalized estimator. This makes it ideally suited for our task, as we expect the generative vector of topic proportions $T^{(i)}$ to be sparse (not all topics are covered by each document) and $\widehat{T}^{(i)}$ can recover the true sparsity pattern, see [21] for details.

Error bounds for the corpora distance: Our proposed estimator of the corpora distance is the plug-in estimator corresponding to the partial estimates explained above. Its error bound is given below, and proved in Appendix C.1. We note that it combines additively the error bounds given in (3.15) and (3.17), obtained under the conditions stated in Corollary 3 of [24] and Corollary 10 in [21]. In particular, we require each corpus to have a balanced number of topics. Formally, for Corpus 1, with $\theta_{\min} := \min_{a \in [K]} \theta_a$ and $\theta_{\max} := \max_{a \in [K]} \theta_a$, we assume that $\theta_{\min} \asymp \theta_{\max}$. It is well understood that it is difficult to estimate A optimally in the absence of this

assumption, see e.g. [7] for an early reference. We make a similar assumption for Corpus 2.

Theorem 26. *If the estimators $\widehat{A}$ and $\widehat{T}^{(i)}$, $i \in [n]$, for Corpus 1 achieve, respectively, the rates (3.15) and (3.17), and if the same rates hold in Corpus 2, then*

$$|W_1(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}'; \widehat{d}^{\text{cluster}}) - W_1(\boldsymbol{\theta}, \boldsymbol{\theta}'; d^{\text{cluster}})| = O_{\mathbb{P}}\Bigg(\sqrt{\frac{\|A\|_0 \log(p)}{nN}} + \sqrt{\frac{\|A'\|_0 \log(p)}{n'N'}} + \theta_{\min}^{-1} \sqrt{\frac{K \log(p)}{N}} + \theta_{\min}'^{-1} \sqrt{\frac{K' \log(p)}{N'}} \Bigg). \tag{3.18}$$

Remark 8. The error rate in estimating the corpus distance cumulates the errors induced by the estimation of its ingredients, $A, A', T^{(i)}, T'^{(i)}$. Furthermore, the quantity $\theta_{\min}$ appears in the rate (3.18) due to the estimation of the within-cluster document weights $\gamma^{(a)} = nT_a^{(i)}/\theta_a$, $a \in [K]$. Since we work under the assumption that the topics are approximately balanced, we have $\theta_{\min} \asymp \theta_{\max} > 1/K$, the latter inequality holding since $\sum_{a=1}^{K} \theta_a = 1$. This shows that the rate of the estimator of our proposed hierarchical Wasserstein distance could include an extra factor of $K \vee K'$ in the term involving cluster center distribution estimates. We suspect that this factor is unavoidable, and defer to the future a careful minimax-rate analysis for the estimation of this distance.

3.4 Application: comparing news sources

In this section we demonstrate the use of our distance with an application to comparing news sources. We compare the New York Times (NYT) to four other news sources by computing the distance between corpora of articles from each source, using a variety of corpus distances. Each corpus consists of documents

from June 2005 from the corresponding news source in the Gigaword English dataset [82].

See Table 3.1 for the results. The news sources are arranged from left to right in order of increasing geographic and cultural dissimilarity in content from the NYT (US, US-based world news, French, Chinese); we expect an informative corpus distance to respect this ordering.

The distances we consider are as follows, from the first row of the table to the last: (i) Our proposed hierarchical distance (HWCD) with word-topic matrix and number of topics estimated using Sparse-TOP [24], and document-topic vectors T estimated by the MLE; and LDA [31] with the same number of topics, as estimated by Sparse-TOP. (ii) We consider a corpus distance (Agg) that does not use the proposed probabilistic representation of each corpus, but instead aggregates an $n \times n'$ matrix D of between-document distances (where n, n' are the number of documents in the two corpora) by computing $\frac{1}{2n}\sum_i \max_j D_{ij} + \frac{1}{2n'}\sum_j \max_i D_{ij}$. For the between-document distances in D we use the topic-level Wasserstein distance from [21]. (iii) Finally, we consider the method of [46], with document-topic matrix estimated using both Sparse-TOP and LDA (with the same number of topics).

Discussion of results: HWCD and Agg both successfully align with our expectations that the distance between the news sources increases (left to right), based on cultural and geographic similarity of the content. However, Table 3.1 shows that Agg has a computation time several orders of magnitude larger than all other methods, limiting its practical utility. Furthermore, HWCD gives rise to an interpretable transport plan between the topic-based cluster centers (see Section 3.2.2), showing how topics are connected between news sources; see

Figure 3.2.

Neither of the distances based on [46] capture the expected relative distances to NYT: Foth.-TOP ranks NYT as closer to XIN and AFP than to LTW, and Foth.-LDA ranks NYT as nearly equidistant to AWP, AFP, and XIN, with AWP being the most distant. Furthermore, these distances are not robust to the topic model estimation method, and can yield contradictory results (0.280 vs. 0.997, for example), in contrast with the Wasserstein-based corpus distances.

3.5 Conclusion

We have defined a new approach to measure the distance between ensembles of independent, but not identically distributed, discrete samples, when each ensemble follows a topic model. This distance, a hierarchical Wasserstein distance on topic-based cluster center distributions, simultaneously captures differences in the content of the topics (as measured by word-topic matrices A) and their relative frequency (measured by weights $(\theta_a)_{a\in[K]}$). We provided a method of estimation of the corpora distance together with theoretical bound on the error of estimation. Finally, we demonstrated its use with an application to newswire data, demonstrating how the distance can be used to detect and interpret topical similarity between news sources.

	News source				
Method	LTW	APW	AFP	XIN	Time (s)
HWCD-TOP	0.202	0.380	0.435	0.519	0.05
HWCD-LDA	0.357	0.500	0.567	0.649	0.03
Agg	0.188	0.349	0.414	0.498	1444.8
Foth.-TOP	0.397	1.000	0.350	0.280	0.03
Foth.-LDA	0.150	1.000	0.988	0.997	0.03

Table 3.1: Distance from the NYT corpus to four other news sources: LA Times/Washington Post (LTW), Associated Press Worldstream (AWP), Agence France-Presse (AFP), and Xinhua News (XIN). The rightmost column gives average computation time.

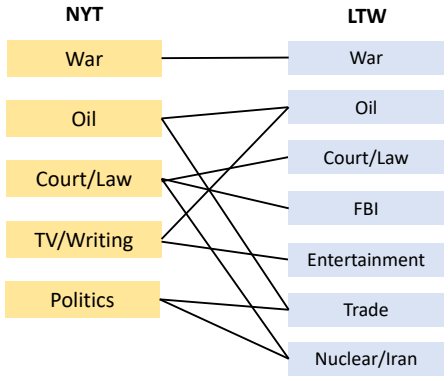

Figure 3.2: The transport plan between NYT and LTW corresponding to the HCWD. We recall that the optimal transport plan w^* is a joint distribution with marginals $\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}'$ that is a solution to the optimization problem (3.13). We draw a line between any topics $k \in [K]$, $k' \in [K']$ with a nonzero value for the transport plan, $w^*_{k,k'} > 0$. The plan depicted here shows how the topical similarity between the NYT and LTW corpora is realized: for example, both sources cover "War", "Oil", and "Court/Law", and "Politics" in the NYT is connected to both "Trade" and "Nuclear/Iran" in LTW.

APPENDIX A

APPENDIX OF CHAPTER 1

A.1 Proofs for Section 1.2

A.1.1 Proof of Theorem 1

We work on the event

$$\mathcal{K} := \left\{\sigma_n^2(\boldsymbol{X}) \gtrsim \operatorname{tr}(\Sigma_X),\ \|\mathbf{y}\|^2 \lesssim n\sigma_y^2\right\}. \tag{A.1}$$

On this event, recalling $\widehat{\alpha} = \boldsymbol{X}^+\mathbf{y}$ and invoking identity (A.101) in Appendix A.5,

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

By Lemma 27 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 (A.2) 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 $\boldsymbol{X} = \tilde{\mathbf{X}}\Sigma_X^{1/2}$, and thus

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

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 28 below thus implies that if $\mathrm{r_e}(\Sigma_X) > C \cdot n$ for $C > 0$ large enough, then with probability at least $1 - 2e^{-cn}$,

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

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

$$\mathbb{P}(\sigma_n^2(\boldsymbol{X}) \gtrsim \mathrm{tr}(\Sigma_X)) \geq 1 - 2e^{-cn}. \tag{A.3}$$

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 [91]) implies that

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

Thus,

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

Combining this with (A.3) establishes that $\mathbb{P}(\mathcal{K}) \geq 1 - ce^{-c'n}$, thus completing the proof. ■

A.1.2 Lemma 27 and Theorem 28

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

Lemma 27. *For any vector $\theta \in \mathbb{R}^p$,*

$$\left|\frac{R(\theta)}{R(\mathbf{0})} - 1\right| \le \frac{\|\theta\|_{\Sigma_X}^2}{R(\mathbf{0})} + 2\sqrt{\frac{\|\theta\|_{\Sigma_X}^2}{R(\mathbf{0})}}. \tag{A.4}$$

Proof. We first show that $\Sigma_X \alpha^* = \Sigma_{XY}$, where $\Sigma_{XY} := \mathbb{E}[Xy]$ and $\alpha^* := \Sigma_X^+ \Sigma_{XY}$. 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 A.5). Thus $(I_p - \Sigma_X \Sigma_X^+)X = 0$ a.s., so

$$\Sigma_X \alpha^* = \Sigma_X \Sigma_X^+ \Sigma_{XY} = \mathbb{E}[\Sigma_X \Sigma_X^+ Xy] = \mathbb{E}[Xy] = \Sigma_{XY}. \tag{A.5}$$

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

$$\begin{aligned}
R(\theta) - R(\mathbf{0}) &= \mathbb{E}[(X^\top \theta - y)^2] - \mathbb{E}[y^2] \\
&= \theta^\top \mathbb{E}[XX^\top]\theta - 2\theta^\top \mathbb{E}[Xy] \\
&= \|\theta\|_{\Sigma_X}^2 - 2\theta^\top \Sigma_{XY} \\
&= \|\theta\|_{\Sigma_X}^2 - 2\theta^\top \Sigma_X \alpha^* \qquad \text{(by (A.5))},
\end{aligned}$$

so by the Cauchy-Schwarz inequality,

$$|R(\theta) - R(\mathbf{0})| \le \|\theta\|_{\Sigma_X}^2 + 2\|\theta\|_{\Sigma_X}\|\alpha^*\|_{\Sigma_X}. \tag{A.6}$$

Next observe that

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

where we use that by (A.5),

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

Thus, $\|\alpha^*\|_{\Sigma_X}^2 \leq R(\mathbf{0})$, so by (A.6),

$$|R(\theta) - R(\mathbf{0})| \leq \|\theta\|_{\Sigma_X}^2 + 2\|\theta\|_{\Sigma_X}\sqrt{R(\mathbf{0})}. \tag{A.7}$$

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

Theorem 28. *Suppose $\boldsymbol{W}$ 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 $1 - 2e^{-cn}$,*

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

where $M := \max_{i,j} \|\boldsymbol{W}_{ij}\|_{\psi_2}$.[1]

A similar result for diagonal Σ has been derived in Lemma 9 of [13]. We make use of the Hanson-Wright inequality in our proof to deal with non-diagonal Σ. Theorem 4.6.1 in [91] provides similar two-sided bounds for the smallest and largest eigenvalue of $\boldsymbol{W}\Sigma\boldsymbol{W}^\top$, when $\Sigma = I_r$.

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

$$\|\boldsymbol{W}\Sigma\boldsymbol{W}^\top - \mathrm{tr}(\Sigma)I_n\| \leq c'(M^2 + M^4)\|\Sigma\|n + \mathrm{tr}(\Sigma)/2 \tag{A.8}$$

with probability at least $1 - 2e^{-cn}$. Equation (A.8) implies that for any $v \in \mathbb{R}^n$ with $\|v\| = 1$,

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

and so

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

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

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

We now prove (A.8). Let $\mathbb{N}$ be a 1/4-net of S^{n-1} with $|\mathbb{N}| \leq 9^n$, which exists by Corollary 4.2.13 of [91]. Then by Exercise 4.4.3 of [91],

$$\|\boldsymbol{W}\Sigma\boldsymbol{W}^\top - \operatorname{tr}(\Sigma)I_n\| = \sup_{v \in S^{n-1}} |v^\top \boldsymbol{W}\Sigma\boldsymbol{W}^\top v - \operatorname{tr}(\Sigma)| \leq 2 \sup_{v \in \mathbb{N}} |v^\top \boldsymbol{W}\Sigma\boldsymbol{W}^\top v - \operatorname{tr}(\Sigma)|, \tag{A.9}$$

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

Now fix $v \in S^{n-1}$ and define $B = \boldsymbol{W}^\top v \in \mathbb{R}^r$. Observe that B has mean zero entries that are independent because the columns of $\boldsymbol{W}$ are independent. Furthermore, by Proposition 2.6.1 of [91],

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

where we used $\|v\|^2 = 1$ in the last step. Thus, by the Hanson-Wright inequality (Theorem 6.2.1 in [91]),

$$\mathbb{P}\left(|B^\top \Sigma B - \mathbb{E}B^\top \Sigma B| \geq c_1 M^2 t\right) \leq 2\exp\left\{-c_2 \min\left(t/\|\Sigma\|, t^2/\|\Sigma\|_F^2\right)\right\}, \tag{A.10}$$

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

Note that

$$\mathbb{E}B^\top \Sigma B = \sum_{i,j,k,l} \mathbb{E}v_i \boldsymbol{W}_{ij} \Sigma_{jl} \boldsymbol{W}_{kl} v_k = \sum_{ij} v_i^2 \Sigma_{jj} \mathbb{E}\boldsymbol{W}_{ij}^2 = \|v\|^2 \operatorname{tr}(\Sigma) = \operatorname{tr}(\Sigma), \tag{A.11}$$

where in the second step we use that $\boldsymbol{W}$ has independent mean zero entries, in the third step we use that $\mathbb{E}\boldsymbol{W}_{ij}^2 = 1$ 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 (A.10) 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,

$$\mathbb{P}\left(|B^\top \Sigma B - \mathrm{tr}(\Sigma)| \geq c_1 M^2 \|\Sigma\| n/2 + c_1 M^2 \sqrt{n\|\Sigma\|_F^2}/2 \right) \leq 2e^{-3n}, \tag{A.12}$$

where we used (A.11). 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 (A.12), and for $c' > 0$ large enough,

$$\mathbb{P}\left(|B^\top \Sigma B - \mathrm{tr}(\Sigma)| \geq c'(M^2 + M^4)\|\Sigma\| n + \mathrm{tr}(\Sigma)/4 \right) \leq 2e^{-3n}. \tag{A.13}$$

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

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{W}\Sigma\boldsymbol{W}^\top - \mathrm{tr}(\Sigma) I_n\| \geq 2L \right) &\leq \mathbb{P}\left(2 \sup_{v \in \mathbb{N}} |v^\top \boldsymbol{W}\Sigma\boldsymbol{W}^\top v - \mathrm{tr}(\Sigma)| \geq 2L \right) && \text{(by (A.9))} \\
&\leq \sum_{v \in \mathbb{N}} \mathbb{P}\left(|v^\top \boldsymbol{W}\Sigma\boldsymbol{W}^\top v - \mathrm{tr}(\Sigma)| \geq L \right) && \text{(union bound)} \\
&\leq 2 \times 9^n e^{-3n} && \text{(by (A.13))} \\
&= 2e^{n \log(9) - 3n} \leq 2e^{-cn},
\end{aligned}$$

where we define $c = 3 - \log(9) > 0$ in the last step. This shows (A.8) and completes the proof. ■

A.2 Proofs for Section 1.3

A.2.1 Proof of Lemma 3 from Section 1.3.1

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

$$\lambda_i(\Sigma_X) = \min_{S:\dim(S)=i} \max_{x\in S:\|x\|=1} x^\top \Sigma_X x, \tag{A.14}$$

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 (A.14), 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 (A.14) 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 (A.14), 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 $\text{rank}(A\Sigma_Z A^\top) \leq K$, 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.

■

A.2.2 Proof of Lemma 4 from Section 1.3.2

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

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

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

$$R(\alpha^*) - \sigma^2 = \mathbb{E}[(\alpha^{*\top}X - Z^\top\beta)]^2 = \alpha^{*\top}\Sigma_X\alpha^* + \beta^\top\Sigma_Z\beta - 2\alpha^{*\top}A\Sigma_Z\beta. \tag{A.15}$$

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 $\bar{A} := A\Sigma_Z^{1/2}$ and $\bar{\beta} := \Sigma_Z^{1/2}\beta$, we have $\alpha^* = \Sigma_X^{-1}\bar{A}\bar{\beta}$. Plugging this into (A.15) and simplifying, we find

$$R(\alpha^*) - \sigma^2 = \bar{\beta}^\top\left[I_K - \bar{A}^\top\Sigma_X^{-1}\bar{A}\right]\bar{\beta}. \tag{A.16}$$

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 (A.16), we find

$$R(\alpha^*) - \sigma^2 = \bar{\beta}^\top \bar{G}^{-1} \bar{\beta} = \bar{\beta}^\top (I_K + \bar{A} \Sigma_E^{-1} \bar{A})^{-1} \bar{\beta}. \tag{A.17}$$

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

$$R(\alpha^*) - \sigma^2 = \bar{\beta}^\top H^{-1/2} (I_K + H^{-1})^{-1} H^{-1/2} \bar{\beta}. \tag{A.18}$$

For the lower bound, first observe that

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

Furthermore,

$$\lambda_K(H) = \lambda_K(\bar{A}^\top \Sigma_E^{-1} \bar{A}) \geq \lambda_K(A \Sigma_Z A^\top) / \|\Sigma_E\| = \xi, \tag{A.19}$$

so using this in the previous display,

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

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

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

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 (A.19) in the last step.

■

A.2.3 Proofs for Section 1.3.3

Proof of Lemma 5

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

$$\alpha^* = \Sigma_X^+ \bar{A}\bar{\beta} = (\bar{A}\bar{A}^\top)^+ \bar{A}\bar{\beta} = \bar{A}^{+\top}\bar{\beta}, \tag{A.20}$$

where we use Lemma 39 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 39.

Next, by identity (A.95) in Lemma 39, 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 (A.20) 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 39. Thus, again using $A^+ A = A^\top A^{+\top} = I_K$, we find

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

as claimed. ■

Proof of Lemma 6

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 A.5). Thus,

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

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

$$\alpha^* = \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{A}^\top\bar{A}^{+\top}\bar{\beta} = \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{\beta}, \tag{A.21}$$

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

Bounds on $\|\alpha^*\|_{\Sigma_X}^2$*:* By (A.21), we find

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

Expanding the above and simplifying, we find

$$\|\alpha^*\|_{\Sigma_X}^2 = \bar{\beta}^\top[I_K - \bar{G}^{-1}]\bar{\beta} = \|\beta\|_{\Sigma_Z}^2 - \bar{\beta}^\top\bar{G}^{-1}\bar{\beta}. \tag{A.22}$$

Recalling that $R(\alpha^*) - \sigma^2 = \bar{\beta}^\top \bar{G}^{-1}\bar{\beta}$ from (A.17) above, Lemma 4 implies that

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

Combining this with (A.22) 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 (A.21), we find

$$\|\alpha^*\|^2 = \bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top \Sigma_E^{-2}\bar{A}\bar{G}^{-1}\bar{\beta}. \tag{A.23}$$

Thus,

$$\begin{aligned}
\|\alpha^*\|^2 &\le \frac{1}{\lambda_p(\Sigma_E)}\bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top \Sigma_E^{-1}\bar{A}\bar{G}^{-1}\bar{\beta} \\
&= \frac{1}{\lambda_p(\Sigma_E)}\bar{\beta}^\top \bar{G}^{-1}(\bar{G} - I_K)\bar{G}^{-1}\bar{\beta} \\
&= \frac{1}{\lambda_p(\Sigma_E)}(\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} - \bar{\beta}^\top \bar{G}^{-2}\bar{\beta}) \\
&\le \frac{1}{\lambda_p(\Sigma_E)}\bar{\beta}^\top \bar{G}^{-1}\bar{\beta}. \qquad \text{(A.24)}
\end{aligned}$$

We also have

$$\begin{aligned}
\|\alpha^*\|^2 &\ge \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}\bar{A}^\top {\Sigma_E}^{-1}\bar{A}\bar{G}^{-1}\bar{\beta} \\
&= \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}(\bar{G} - I_K)\bar{G}^{-1}\bar{\beta} \\
&= \frac{1}{\|\Sigma_E\|}[\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} - \bar{\beta}^\top \bar{G}^{-2}\bar{\beta}] \\
&\ge \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} \cdot [1 - 1/\lambda_K(\bar{G})] \qquad \text{(A.25)} \\
&\ge \frac{1}{\|\Sigma_E\|}\bar{\beta}^\top \bar{G}^{-1}\bar{\beta} \cdot [1 - 1/\xi], \qquad \text{(A.26)}
\end{aligned}$$

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 (A.24) and (A.26),

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

Recalling that $R(\alpha^*) - \sigma^2 = \bar{\beta}^\top \bar{G}^{-1} \bar{\beta}$ from (A.17) above, Lemma 4 implies

$$\left(\frac{\xi - 1}{\xi + 1}\right) \frac{1}{\|\Sigma_E\|} \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \leq \|\alpha^*\|^2 \leq \frac{1}{\lambda_p(\Sigma_E)} \beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta. \tag{A.27}$$

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 (A.27) proves that

$$\left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \cdot \beta^\top (A^\top A)^{-1} \beta \leq \|\alpha^*\|^2 \leq \kappa(\Sigma_E) \cdot \beta^\top (A^\top A)^{-1} \beta. \tag{A.28}$$

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

Proof of (A.28): 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 rank$(A) = K$). Then,

$$(A^\top \Sigma_E^{-1} A)^{-1} = (V_A S_A U_A^\top \Sigma_E^{-1} U_A S_A V_A^\top)^{-1} = V_A S_A^{-1} (U_A^\top \Sigma_E^{-1} U_A)^{-1} S_A^{-1} V_A^\top. \tag{A.29}$$

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

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

we find

$$\beta^\top (A^\top \Sigma_E^{-1} A)^{-1} \beta \ge \lambda_p(\Sigma_E) \cdot \beta^\top V_A S_A^{-2} V_A^\top \beta. \tag{A.30}$$

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

$$(A^\top A)^{-1} = (V_A S_A U_A^\top U_A S_A V_A^\top)^{-1} = V_A S_A^{-2} V_A^\top, \tag{A.31}$$

and thus, by (A.30),

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

which proves the lower bound in (A.28). To prove the upper bound, we use that by (A.29),

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

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 (A.31). This establishes the upper bound of (A.28), completing the proof.

■

Proof of Corollary 7

Under the conditions stated, by either Lemma 5 or Lemma 6, $\|\alpha^*\|^2 \lesssim \beta^\top (A^\top A)^{-1}\beta$. Thus, using that Σ_Z is invertible,

$$\|\alpha^*\|^2 \lesssim \beta^\top (A^\top A)^{-1}\beta = \beta^\top \Sigma_Z^{1/2}(\Sigma_Z^{1/2} A^\top A \Sigma_Z^{1/2})^{-1}\Sigma_Z^{1/2}\beta \le \|\beta\|^2_{\Sigma_Z}/\lambda_K(A\Sigma_Z A^\top), \tag{A.32}$$

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

For the second claim, we have

$$\begin{aligned} R(\mathbf{0}) - R(\alpha^*) &= \|\alpha^*\|^2_{\Sigma_X} && \text{(by (1.16))} \\ &\gtrsim \|\beta\|^2_{\Sigma_Z}. && \text{(by either Lemma 5 or Lemma 6)} \end{aligned}$$

The claim follows by taking the limit inferior as $p \to \infty$ on both sides of the inequality and using condition (1.15). ■

A.3 Proofs for Section 1.4

A.3.1 Proofs for Section 1.4.1

In the proofs of Lemma 8 and Theorem 9, we will use the event

$$\mathcal{A} := \left\{ \|\tilde{\mathbf{Z}}^+ \tilde{\boldsymbol{\varepsilon}}\|^2 \lesssim \log(n)\mathrm{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^+),\ c_1 n \le \sigma_K^2(\tilde{\mathbf{Z}}) \le \|\tilde{\mathbf{Z}}\|^2 \le c_2 n \right\}, \tag{A.33}$$

which occurs with probability at least $1 - c/n$, as shown in Lemma 29 below, where $\boldsymbol{Z} = \tilde{\mathbf{Z}}\Sigma_Z^{1/2}$ and $\boldsymbol{\varepsilon} = \sigma\tilde{\boldsymbol{\varepsilon}}$ by Assumption 3.

Proof of Lemma 8

On the event $\mathcal{A}$ defined in (A.33), and using $\lambda_K(\Sigma_Z) > 0$ by Assumption 2,

$$\sigma_K^2(\boldsymbol{Z}) = \lambda_K(\boldsymbol{Z}\boldsymbol{Z}^\top) = \lambda_K(\tilde{\boldsymbol{Z}}\Sigma_Z\tilde{\boldsymbol{Z}}^\top) \geq \lambda_K(\Sigma_Z) \cdot \sigma_n^2(\tilde{\boldsymbol{Z}}) \gtrsim \lambda_K(\Sigma_Z) \cdot n > 0, \tag{A.34}$$

so $\text{rank}(\boldsymbol{Z}) = K$ and thus $\boldsymbol{Z}^+\boldsymbol{Z} = I_K$ by Lemma 39 in Appendix A.5. 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 (A.95) of Lemma 39, we find

$$\boldsymbol{X}^+ = (\boldsymbol{Z}A^\top)^+ = (\boldsymbol{Z}^+\boldsymbol{Z}A^\top)^+(\boldsymbol{Z}A^\top A^{+\top})^+ = A^{+\top}\boldsymbol{Z}^+. \tag{A.35}$$

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

$$\widehat{\alpha} = \boldsymbol{X}^+\mathbf{y} = A^{+\top}\boldsymbol{Z}^+\mathbf{y}, \tag{A.36}$$

so

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

where in the last step we used that by Lemma 39,

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

Continuing, and using

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

we find

$$\begin{aligned}
\|\widehat{\alpha}\|^2 &\lesssim \beta^\top (A^\top A)^{-1}\beta + \|(A\Sigma_Z^{1/2})^+\|^2 \cdot \sigma^2 \cdot \|\tilde{\mathbf{Z}}^+\tilde{\boldsymbol{\varepsilon}}\|^2 \\
&\lesssim \beta^\top (A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma^2 \log(n)\mathrm{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^+) && \text{(on } \mathcal{A}) \\
&\leq \beta^\top (A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma^2 \log(n) K \|\tilde{\mathbf{Z}}^+\|^2 \\
&= \beta^\top (A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma^2 \log(n) K \frac{1}{\sigma_K^2(\tilde{\mathbf{Z}})} \\
&\lesssim \beta^\top (A^\top A)^{-1}\beta + \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\sigma^2 \log(n)\frac{K}{n} && \text{(on } \mathcal{A}) \\
&\leq \frac{1}{\lambda_K(A\Sigma_Z A^\top)}\left(\|\beta\|_{\Sigma_Z}^2 + \sigma^2 \log(n)\frac{K}{n}\right). && \text{(by (A.32))}
\end{aligned}$$

Under the assumptions of this Lemma, the event $\mathcal{A}$ holds with probability at least $1 - c/n$ by Lemma 29, so the proof is complete. ■

Proof of Theorem 9

Part 1: By (A.36), $\widehat{\alpha} = A^{+\top}\mathbf{Z}^+\mathbf{y}$ on the event $\mathcal{A}$ defined in (A.33). Thus, using $X = AZ$ and $A^\top A^{+\top} = I_K$ since A is full rank by Assumption 2,

$$\widehat{y}_x = X^\top\widehat{\alpha} = Z^\top A^\top A^{+\top}\mathbf{Z}^+\mathbf{y} = Z^\top\mathbf{Z}^+\mathbf{y} = Z^\top\widehat{\beta} = \widehat{y}_z. \tag{A.37}$$

Part 2: Using the independence of ε and Z together with (A.37), the excess risk can be written as

$$R(\widehat{\alpha}) - \sigma^2 = \mathbb{E}[(X^\top\widehat{\alpha} - Z^\top\beta)^2] = \mathbb{E}[(Z^\top\widehat{\beta} - Z^\top\beta)^2] = \|\widehat{\beta} - \beta\|_{\Sigma_Z}^2. \tag{A.38}$$

By (A.34), $\mathrm{rank}(\mathbf{Z}) = K$ and $\mathbf{Z}^+\mathbf{Z} = I_K$ on the event $\mathcal{A}$ defined in (A.33). Thus,

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

so by (A.38),

$$R(\widehat{\alpha}) - \sigma^2 = \|\mathbf{Z}^+\boldsymbol{\varepsilon}\|_{\Sigma_Z}^2 = \|\Sigma_Z^{1/2}\mathbf{Z}^+\boldsymbol{\varepsilon}\|^2. \tag{A.39}$$

By (A.95) of Lemma 39,

$$\Sigma_Z^{1/2}\mathbf{Z}^+ = \Sigma_Z^{1/2}(\tilde{\mathbf{Z}}\Sigma_Z^{1/2})^+ = \Sigma_Z^{1/2}(\tilde{\mathbf{Z}}^+\tilde{\mathbf{Z}}\Sigma_Z^{1/2})^+(\tilde{\mathbf{Z}}\Sigma_Z^{1/2}\Sigma_Z^{-1/2})^+ = \Sigma_Z^{1/2}\Sigma_Z^{-1/2}\tilde{\mathbf{Z}}^+ = \tilde{\mathbf{Z}}^+, \quad \text{(A.40)}$$

where we used that $\tilde{\mathbf{Z}}^+\tilde{\mathbf{Z}} = I_K$ since $\text{rank}(\tilde{\mathbf{Z}}) = K$ on $\mathcal{A}$. Thus by (A.39), we find that on $\mathcal{A}$,

$$R(\widehat{\alpha}) - \sigma^2 = \|\tilde{\mathbf{Z}}^+\boldsymbol{\varepsilon}\|^2 = \sigma^2\|\tilde{\mathbf{Z}}^+\tilde{\boldsymbol{\varepsilon}}\|^2 \lesssim \sigma^2 \log(n)\text{tr}(\tilde{\mathbf{Z}}^{+\top}\tilde{\mathbf{Z}}^+). \quad \text{(A.41)}$$

We then use that $\text{rank}(\tilde{\mathbf{Z}}^+) = K$ and that $\|\tilde{\mathbf{Z}}^+\| = 1/\sigma_K(\tilde{\mathbf{Z}})$ from Lemma 39 in Appendix A.5 below to find that on $\mathcal{A}$,

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

Plugging this into (A.41) completes the proof of the upper bound.

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

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

so using $\sigma_K(\tilde{\mathbf{Z}}^+) = 1/\|\tilde{\mathbf{Z}}\|$ by Lemma 39 again,

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

■

Lemma 29. *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*

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

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 [91] gives that with probability at least $1 - 2/n$,

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

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

$$\mathbb{P}\left(c_3 n \leq \sigma_K^2(\tilde{\mathbf{Z}}) \leq \|\tilde{\mathbf{Z}}\|^2 \leq c_4 n\right) \geq 1 - 2/n. \tag{A.42}$$

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 30, which we state below. Combining this with (A.42) proves that $\mathcal{A}$ occurs with probability at least $1 - c/n$. ■

The following result is a slightly adapted version of Lemma 19 from [13] and the discussion that follows.

Lemma 30. *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 $1 - e^{-cn}$,*

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

Proof of Lemma 10

Suppose $\mathrm{rank}(\boldsymbol{X}) = K$. We can then write the singular value decomposition of $\boldsymbol{X}$ as $\boldsymbol{X} = \widehat{V}_K \widehat{D} \widehat{U}_K^\top$, 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,

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

By Lemma 39 of Appendix A.5, we thus have

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

We thus find

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

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

Now suppose that Assumptions 2 & 3 hold and $K > C \cdot n$. Then by Lemma 29 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,

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

which completes the proof. ■

A.3.2 Proofs for Section 1.4.2

In this section we begin with the proof of Lemma 12 and our main result, Theorem 13, which rely on Proposition 11, proved subsequently. The proofs of Lemma 12 and Theorem 13 use the event

$$\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3, \tag{A.43}$$

where for positive absolute constants c_1 to c_6,

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

$$\mathcal{E}_2 := \{\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{X}^{+\top} \Sigma_X \boldsymbol{X}^+ \tilde{\boldsymbol{\varepsilon}} \leq c_5 \log(n) \mathrm{tr}(\boldsymbol{X}^{+\top} \Sigma_X \boldsymbol{X}^+)\},$$

$$\mathcal{E}_3 := \{\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{X}^{+\top} \boldsymbol{X}^+ \tilde{\boldsymbol{\varepsilon}} \leq c_6 \log(n) \mathrm{tr}(\boldsymbol{X}^{+\top} \boldsymbol{X}^+)\}.$$

We will show in Lemma 31 below that $\mathcal{E}$ occurs with probability at least $1 - c/n$ for an absolute constant $c > 0$.

Proof of Theorem 12

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

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

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

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

where in the last step holds on the event $\mathcal{E}$, and uses that $\|\boldsymbol{X}^+ \boldsymbol{X}\| \leq 1$ since $\boldsymbol{X}^+ \boldsymbol{X}$ is a projection matrix. Recalling that by (A.32),

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

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

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

By Lemma 31, $\mathcal{E}$ holds with probability at least $1 - c/n$, so the proof is complete.■

Proof of Theorem 13

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

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

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

$$\|\Sigma_E^{1/2}\widehat{\alpha}\|^2 \le 2\|\Sigma_E^{1/2}\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2 + 2\|\Sigma_E^{1/2}\boldsymbol{X}^+\boldsymbol{\varepsilon}\|^2 := 2B_1 + 2V_1.$$

Similarly,

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

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

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

Bounding the bias component: On the event $\mathcal{E}$ defined in (A.43), $\sigma_n(\boldsymbol{X}) > 0$ and by Assumption 2 and (A.34) above, $\sigma_n^2(\boldsymbol{Z}) \gtrsim \lambda_K(\Sigma_Z)n > 0$. Thus $\boldsymbol{X}$ and $\boldsymbol{Z}$ are of rank n and K respectively, so by Lemma 39 of Appendix A.5, $\boldsymbol{X}\boldsymbol{X}^+ = I_n$ and $\boldsymbol{Z}^+\boldsymbol{Z} = I_K$. It follows that

$$\begin{aligned}
\boldsymbol{Z}^+ - A^\top \boldsymbol{X}^+ &= \boldsymbol{Z}^+\boldsymbol{X}\boldsymbol{X}^+ - A^\top \boldsymbol{X}^+ && (\text{since } \boldsymbol{X}\boldsymbol{X}^+ = I_n) \\
&= (\boldsymbol{Z}^+\boldsymbol{X} - A^\top)\boldsymbol{X}^+ \\
&= (\boldsymbol{Z}^+[\boldsymbol{Z}A^\top + \mathcal{E}] - A^\top)\boldsymbol{X}^+ && (\text{since } \boldsymbol{X} = \boldsymbol{Z}A^\top + \mathcal{E}) \\
&= \boldsymbol{Z}^+\mathcal{E}\boldsymbol{X}^+, && (\text{since } \boldsymbol{Z}^+\boldsymbol{Z} = I_K) \qquad \text{(A.44)}
\end{aligned}$$

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

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

By (A.40) above and the fact that $\boldsymbol{Z}$ is full rank on $\mathcal{E}$, $\Sigma_Z^{1/2}\boldsymbol{Z}^+ = \tilde{\boldsymbol{Z}}^+$, so on $\mathcal{E}$,

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

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

$$B = B_1 + B_2 \lesssim \frac{\mathrm{tr}(\Sigma_E)\|\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2}{n}. \qquad \text{(A.45)}$$

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

$$\|\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2 = \|\boldsymbol{X}^+\boldsymbol{Z}A^\top A^{+\top}\beta\|^2 \le 2\|\boldsymbol{X}^+\boldsymbol{X}A^{+\top}\beta\|^2 + 2\|\boldsymbol{X}^+\mathcal{E}A^{+\top}\beta\|^2.$$

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

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

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

$$\|\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2 \lesssim \|A^{+\top}\beta\|^2. \tag{A.46}$$

Finally, we have

$$\|A^{+\top}\beta\|^2 = \beta^\top(A^\top A)^{-1}\beta = \beta^\top\Sigma_Z^{1/2}(\Sigma_Z^{1/2}A^\top A\Sigma_Z^{1/2})^{-1}\Sigma_Z^{1/2}\beta \le \frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)}. \tag{A.47}$$

Combining this with (A.46) and plugging into (A.45), we find that on the event $\mathcal{E}$,

$$B \lesssim \frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)}\frac{\mathrm{tr}(\Sigma_E)}{n} = \frac{\|\beta\|_{\Sigma_Z}^2\|\Sigma_E\|}{\lambda_K(A\Sigma_Z A^\top)}\cdot\frac{\mathrm{tr}(\Sigma_E)}{\|\Sigma_E\|n} = \frac{\|\beta\|_{\Sigma_Z}^2}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}. \tag{A.48}$$

Bounding the variance component: First note that

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

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

$$V \lesssim \sigma^2\log(n)\mathrm{tr}(\boldsymbol{X}^{+\top}\Sigma_X\boldsymbol{X}^+) = \sigma^2\log(n)\left\{\mathrm{tr}(\boldsymbol{X}^{+\top}\Sigma_E\boldsymbol{X}^+) + \mathrm{tr}(\boldsymbol{X}^{+\top}A\Sigma_Z A^\top\boldsymbol{X}^+)\right\}, \tag{A.49}$$

where we use $\Sigma_X = A\Sigma_Z A^\top + \Sigma_E$ in the second step. The first term in (A.49) can by bounded as

$$\mathrm{tr}(\boldsymbol{X}^{+\top}\Sigma_E\boldsymbol{X}^+) \le \|\Sigma_E\|\cdot n\|\boldsymbol{X}^{+\top}\boldsymbol{X}^+\| = \|\Sigma_E\|\frac{n}{\sigma_n^2(\boldsymbol{X})} \lesssim \frac{n}{\mathrm{r_e}(\Sigma_E)}, \tag{A.50}$$

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

For the second term in (A.49),

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{X}^{+\top}A\Sigma_Z A^\top\boldsymbol{X}^+) &\le K\|\Sigma_Z^{1/2}A^\top\boldsymbol{X}^+\|^2 && \text{(since } \mathrm{rank}(A\Sigma_Z A^\top) = K\text{)}\\
&= K\|\Sigma_Z^{1/2}(\boldsymbol{Z}^+ - \boldsymbol{Z}^+\mathcal{E}\boldsymbol{X}^+)\|^2 && \text{(by (A.44) above)}\\
&\le 2K\|\tilde{\boldsymbol{Z}}^+\|^2 + 2K\|\tilde{\boldsymbol{Z}}^+\|^2\|\mathcal{E}\|^2\|\boldsymbol{X}^+\|^2,
\end{aligned}$$

where we use that $\Sigma_Z^{1/2}\boldsymbol{Z}^+ = \tilde{\boldsymbol{Z}}^+$ from (A.40) in the final step. Continuing, we find

$$\mathrm{tr}(\boldsymbol{X}^{+\top}A\Sigma_Z A^\top \boldsymbol{X}^+) \lesssim \frac{K}{\sigma_K^2(\tilde{\boldsymbol{Z}})}\left(1 + \frac{\|\mathcal{E}\|^2}{\sigma_n^2(\boldsymbol{X})}\right) \lesssim \frac{K}{n}, \tag{A.51}$$

where we use the bounds defining $\mathcal{E}_1$ in the last inequality. Combining (A.51) and (A.50) with (A.49), we conclude that on $\mathcal{E}$,

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

Combining this with the bias bound (A.48) gives the bound in the statement of the theorem. By Lemma 31 below, $\mathbb{P}(\mathcal{E}) \geq 1 - c/n$, so the proof is complete. ■

Lemma 31. *Under model (1.5), 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 $\mathbb{P}(\mathcal{E}) \geq 1 - c/n$, where $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and*

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

$$\mathcal{E}_2 := \{\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{X}^{+\top}\Sigma_X \boldsymbol{X}^+\tilde{\boldsymbol{\varepsilon}} \leq c_5 \log(n) tr(\boldsymbol{X}^{+\top}\Sigma_X\boldsymbol{X}^+)\},$$

$$\mathcal{E}_3 := \{\tilde{\boldsymbol{\varepsilon}}^\top \boldsymbol{X}^{+\top}\boldsymbol{X}^+\tilde{\boldsymbol{\varepsilon}} \leq c_6 \log(n) tr(\boldsymbol{X}^{+\top}\boldsymbol{X}^+)\},$$

for positive constants c_1 to c_6.

Proof. We have $\mathbb{P}(\mathcal{E}^c) \leq \mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c) + \mathbb{P}(\mathcal{E}_3^c)$. The bounds $\mathbb{P}(\mathcal{E}_2^c) \leq e^{-cn}$ and $\mathbb{P}(\mathcal{E}_3^c) \leq e^{-cn}$ follow immediately from Lemma 30 in Appendix A.3.1 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

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

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 11 can be applied to conclude

$$\mathbb{P}\{\sigma_n^2(\boldsymbol{X}) \leq c_1 \mathrm{tr}(\Sigma_E)\} \leq 2e^{-cn}.$$

2. By Assumption 3, $\mathcal{E} = \tilde{\mathbf{E}}\Sigma_E^{1/2}$, where $\tilde{\mathbf{E}}$ has independent entries with zero mean, unit variance, and sub-Gaussian constants bounded by an absolute constant. Thus,

$$\|\mathcal{E}\|^2 = \|\mathcal{E}\mathcal{E}^\top\| = \|\tilde{\mathbf{E}}\Sigma_E\tilde{\mathbf{E}}^\top\|,$$

and by applying Theorem 28 with $\tilde{\mathbf{E}}$ and Σ_E we find that with probability at least $1 - 2e^{-cn}$,

$$\|\mathcal{E}\|^2 \leq \mathrm{tr}(\Sigma_E) + c'\|\Sigma_E\|n = \mathrm{tr}(\Sigma_E) \cdot (1 + c'n/\mathrm{r_e}(\Sigma_E)) \lesssim \mathrm{tr}(\Sigma_E),$$

where the last inequality holds since $n/\mathrm{r_e}(\Sigma_E) < 1/C$. Thus for $c_2 > 0$,

$$\mathbb{P}\{\|\mathcal{E}\|^2 \geq c_2 \mathrm{tr}(\Sigma_E)\} \leq 2e^{-cn}.$$

3. By (A.42) we have that with probability at least $1 - 2/n$,

$$c_3 n \leq \sigma_K^2(\tilde{\mathbf{Z}}) \leq \|\tilde{\mathbf{Z}}\|^2 \leq c_4 n.$$

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

Proof of Proposition 11

We will work on the event

$$\mathcal{F} := \{\sigma_n^2(\mathcal{E}U_{(K+1):p}) \geq c_4 \mathrm{tr}(\Sigma_E),\ \|\tilde{\mathbf{Z}}\|^2 \leq c_5 n\},$$

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, $\mathcal{E} = \tilde{\mathbf{E}}\Sigma_E^{1/2}$, 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

$$\sigma_n^2(\mathcal{E}U_{(K+1):p}) = \lambda_n(\mathcal{E}Q\mathcal{E}^\top) = \lambda_n(\tilde{\mathbf{E}}\Sigma_E^{1/2}Q\Sigma_E^{1/2}\tilde{\mathbf{E}}^\top).$$

We can now apply Theorem 28, stated and proved above in Section A.1, 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 $1 - 2e^{-cn}$,

$$\sigma_n^2(\mathcal{E}U_{(K+1):p}) \geq \operatorname{tr}(\Sigma_E^{1/2}Q\Sigma_E^{1/2})/2 - c'\|\Sigma_E^{1/2}Q\Sigma_E^{1/2}\|n. \tag{A.52}$$

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

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

Plugging these two results into (A.52), we find that with probability at least $1 - 2e^{-cn}$,

$$\sigma_n^2(\mathcal{E}U_{(K+1):p}) \geq \operatorname{tr}(\Sigma_E)/2 - (1/2 + c')n\|\Sigma_E\| = \operatorname{tr}(\Sigma_E)\cdot[1/2 - (1/2+c')n/\mathrm{r_e}(\Sigma_E)] \gtrsim \operatorname{tr}(\Sigma_E), \tag{A.53}$$

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 [91],

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

with probability at least $1 - e^{-c'n}$. Combining this with A.53 we conclude that

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

Bounding $\sigma_n(\boldsymbol{X})$ on $\mathcal{F}$: We now show that $\sigma_n^2(\boldsymbol{X}) \gtrsim \mathrm{tr}(\Sigma_E)$ holds on the event $\mathcal{F}$. Let $\Sigma_X = UDU^\top$ with $U \in \mathbb{R}^{p\times p}$ orthogonal and $D = \mathrm{diag}(\lambda_1(\Sigma_X), \ldots, \lambda_p(\Sigma_X))$. 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

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

so

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

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

$$\sigma_n(\boldsymbol{X}) \geq \sigma_n(\boldsymbol{X}U_{(K+1):p}). \tag{A.54}$$

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

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

so by (A.54),

$$\sigma_n(\boldsymbol{X}) \geq \sigma_n(\boldsymbol{X}U_{(K+1):p}) \geq \sigma_n(\mathcal{E}U_{(K+1):p}) - \|\boldsymbol{Z}A^\top U_{(K+1):p}\| \gtrsim \sqrt{\mathrm{tr}(\Sigma_E)} - \|\boldsymbol{Z}A^\top U_{(K+1):p}\|, \tag{A.55}$$

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

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

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 $\mathcal{F}$,

$$\|\mathbf{Z}A^\top U_{(K+1):p}\|^2 = \|\tilde{\mathbf{Z}}\Sigma_Z^{1/2}A^\top U_{(K+1):p}\| \le \|\tilde{\mathbf{Z}}\|^2\|\Sigma_Z^{1/2}A^\top U_{(K+1):p}\|^2 \lesssim n\|\Sigma_Z^{1/2}A^\top U_{(K+1):p}\|^2. \tag{A.56}$$

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 $D_{(K+1):p} := \mathrm{diag}(\lambda_{K+1}(\Sigma_X), \dots, \lambda_p(\Sigma_X))$,

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

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 (A.56), we find that on $\mathcal{F}$,

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

■

A.3.3 Proof of Theorem 15 from Section 1.4.4

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

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

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 $\boldsymbol{\eta} = \mathbf{y} - \boldsymbol{X}\alpha^*$. We work on the event $\Delta := \Delta_1 \cap \Delta_2$, where

$$\Delta_1 := \left\{ \sigma_K^2(\boldsymbol{X}U_K D_K^{-1/2}) \gtrsim n,\ \|\boldsymbol{X}\Sigma_X^{-1/2}\|^2 \lesssim p \right\},$$

and

$$\Delta_2 := \left\{ \|(\boldsymbol{X}U_K D_K^{-1/2})^+ \boldsymbol{\eta}\|^2 \lesssim \log(n) \cdot \sigma_\eta^2 \cdot \mathrm{tr}[(\boldsymbol{X}U_K D_K^{-1/2})^{+\top} (\boldsymbol{X}U_K D_K^{-1/2})^+] \right\}.$$

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

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

$$\mathbb{E}[X\eta] = \mathbb{E}[Xy] - \mathbb{E}[XX^\top]\alpha^* = \Sigma_{XY} - \Sigma_X \Sigma_X^+ \Sigma_{XY} = 0, \tag{A.57}$$

where we used (A.5) in the last step. Thus,

$$\begin{aligned} R(\widetilde{\alpha}_{\mathrm{PCR}}) &:= \mathbb{E}[(X^\top \widetilde{\alpha}_{\mathrm{PCR}} - y)^2] \\ &= \mathbb{E}\left[(X^\top \widetilde{\alpha}_{\mathrm{PCR}} - X^\top \alpha^* - \eta)^2\right] \\ &= \mathbb{E}\left[(X^\top \widetilde{\alpha}_{\mathrm{PCR}} - X^\top \alpha^*)^2\right] + \mathbb{E}[\eta^2] \qquad \text{(by A.57)} \\ &= \|\widetilde{\alpha}_{\mathrm{PCR}} - \alpha^*\|_{\Sigma_X}^2 + R(\alpha^*). \end{aligned} \tag{A.58}$$

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

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

we find

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

From the fact that $\boldsymbol{X}U_K$ is an $n \times K$ matrix with $K < n$ and $\mathrm{rank}(\boldsymbol{X}U_K) = K$ on the event Δ_1, we have $(\boldsymbol{X}U_K)^+\boldsymbol{X}U_K = I_K$ by Lemma 39 of Appendix A.5 below. Thus, using $P = U_K U_K^\top$ we have $(\boldsymbol{X}U_K)^+\boldsymbol{X}P = U_K^\top$. Applying this in the previous display, we find

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

It thus follows from the decomposition (A.58) that

$$\begin{aligned}R(\widetilde{\alpha}_{\mathrm{PCR}}) - R(\alpha^*) &= \|\widetilde{\alpha}_{\mathrm{PCR}} - \alpha^*\|_{\Sigma_X}^2 \\ &\lesssim \|(I_p - P)\alpha^*\|_{\Sigma_X}^2 + \|U_K(\boldsymbol{X}U_K)^+\boldsymbol{X}(I_p - P)\alpha^*\|_{\Sigma_X}^2 + \|U_K(\boldsymbol{X}U_K)^+\boldsymbol{\eta}\|_{\Sigma_X}^2 \\ &=: B_1 + B_2 + V. \end{aligned} \tag{A.59}$$

Bounding B_1: We find

$$B_1 = \|\Sigma_X^{1/2}(I_p - P)\alpha^*\|^2 \le \|\Sigma_X^{1/2}(I_p - P)\|^2\|\alpha^*\|^2 = \|(I - P)\Sigma_X(I - P)\|\|\alpha^*\|^2. \tag{A.60}$$

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,

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

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

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

and combining this with (A.60) we find

$$B_1 \le \|\Sigma_E\|\|\alpha^*\|^2. \tag{A.61}$$

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

$$\begin{aligned} B_2 &= \alpha^{*\top}(I_p - P)\boldsymbol{X}^\top (\boldsymbol{X}U_K)^{+\top} U_K^\top \Sigma_X U_K (\boldsymbol{X}U_K)^+ \boldsymbol{X}(I - P)\alpha^* \\ &= \|D_K^{1/2}(\boldsymbol{X}U_K)^+ \boldsymbol{X}(I_p - P)\alpha^*\|^2. \end{aligned} \tag{A.62}$$

Observe that by Lemma 39 of Appendix A.5,

$$(\boldsymbol{X}U_K D_K^{-1/2})^+ = [(\boldsymbol{X}U_K)^+ (\boldsymbol{X}U_K) D_K^{-1/2}]^+ \cdot [\boldsymbol{X}U_K D_K^{-1/2} D_K^{1/2}]^+ = D_K^{1/2}(\boldsymbol{X}U_K)^+, \tag{A.63}$$

where we used that $\boldsymbol{X}U_K$ is a full rank $n \times K$ matrix with $K < n$ so $(\boldsymbol{X}U_K)^+(\boldsymbol{X}U_K) = I_K$. Using this in (A.62) yields

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

where the last step holds on Δ. Recalling that $\|\Sigma_X^{1/2}(I_p - P)\alpha^*\|^2 = B_1$ and using (A.61), we find that

$$B_2 \lesssim \|\Sigma_E\| \cdot \|\alpha^*\|^2 \frac{p}{n}. \tag{A.64}$$

Bounding V: We have on Δ,

$$\begin{aligned}
V &= \boldsymbol{\eta}^\top (\boldsymbol{X}U_K)^{+\top} U_K^\top \Sigma_X U_K (\boldsymbol{X}U_K)^+ \boldsymbol{\eta} \\
&= \boldsymbol{\eta}^\top (\boldsymbol{X}U_K)^{+\top} D_K (\boldsymbol{X}U_K)^+ \boldsymbol{\eta} \\
&= \|D_K^{1/2} (\boldsymbol{X}U_K)^+ \boldsymbol{\eta}\|^2 \\
&= \|(\boldsymbol{X}U_K D_K^{-1/2})^+ \boldsymbol{\eta}\|^2 && \text{(by (A.63))} \\
&\lesssim \sigma_\eta^2 \cdot \log(n) \cdot \mathrm{tr}[(\boldsymbol{X}U_K D^{-1/2})^{+\top} (\boldsymbol{X}U_K D^{-1/2})^+] && \text{(on } \Delta_2\text{)} \\
&\le \sigma_\eta^2 \cdot \log(n) \cdot K \cdot \|(\boldsymbol{X}U_K D^{-1/2})^+\|^2 && \text{(since } \mathrm{rank}(\boldsymbol{X}U_K D^{-1/2}) = K\text{)} \\
&= \sigma_\eta^2 \cdot \frac{K \log n}{\sigma_K^2(\boldsymbol{X}U_K D^{-1/2})} \\
&\lesssim \sigma_\eta^2 \cdot \frac{K \log n}{n}. && \text{(on } \Delta_1\text{).}
\end{aligned}$$

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

$$V \lesssim R(\alpha^*) \cdot \frac{K \log n}{n}. \tag{A.65}$$

Combining this with (A.61) and (A.64) proves (1.33).

In the case $\Sigma_E = 0$, the bound (1.34) follows immediately from (1.33). When $\lambda_p(\Sigma_E) > 0$, Lemma 4 of Section 1.3.2 implies

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

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 (1.33) gives

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

where in the second step we use that

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

This proves (1.35). All that remains is to bound the probability of the event Δ.

Bounding $\mathbb{P}(\Delta)$*:* We first bound the probability $\mathbb{P}(\Delta_1)$. Note that the matrix $\boldsymbol{X} U_K D_K^{-1/2}$ 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 $\boldsymbol{X} U_K D_K^{-1/2}$ i.i.d. $N(0,1)$ entries. Thus, by Theorem 4.6.1 of [91], with probability at least $1 - 2/n$,

$$\sigma_K(\boldsymbol{X} U_K D_K^{-1/2}) \geq \sqrt{n} - c(\sqrt{K} + \sqrt{\log n}) = \sqrt{n} \cdot [1 - c\sqrt{K/n} - c\sqrt{\log(n)/n}] \gtrsim \sqrt{n}, \tag{A.66}$$

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

Similarly, $\boldsymbol{X} \Sigma_X^{-1/2}$ is a $n \times p$ matrix with i.i.d. $N(0,1)$ entries, so again by by Theorem 4.6.1 of [91], with probability at least $1 - 2e^{-n}$,

$$\|\boldsymbol{X} \Sigma_X^{-1/2}\| \leq \sqrt{n} + c(\sqrt{p} + \sqrt{n}) \lesssim \sqrt{p}. \tag{A.67}$$

Using a union bound to combine this with (A.66), we find

$$\mathbb{P}(\Delta_1) \geq 1 - c'/n,$$

for some $c' > 0$.

To bound $\mathbb{P}(\Delta_2)$, first note that by (A.57) and the assumption that (X, y) are Gaussian, $\boldsymbol{X}$ and $\boldsymbol{\eta}$ are independent. Furthermore, $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}/\sigma_\eta$ has independent $N(0,1)$ entries. We can thus apply Lemma 30 from Appendix A.3.1 above with

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

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

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

and so $\mathbb{P}(\Delta_2^c) \leq e^{-cn}$. ■

A.3.4 Detailed Comparison of the Bias and Variance Terms in Section 1.4.3

In this sections we give a detailed comparison between our Theorem 13 and Theorem 4 in [13]. 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

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

The bound of [13] 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 (1.5) holds, (X, y) are jointly Gaussian, $\mathrm{rank}(\Sigma_X) \geq n$, and n is large enough, with probability at least $1 - e^{-c'n}$,

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

where

$$B := \|\alpha^*\|^2 \|\Sigma_X\| \max\left\{\sqrt{\frac{r_0(\Sigma_X)}{n}}, \frac{r_0(\Sigma_X)}{n}, 1\right\}, \tag{A.68}$$

and

$$\boldsymbol{V} := \sigma^2 \log(n) \left(\frac{n}{R_{K^*}(\Sigma_X)} + \frac{K^*}{n}\right) \tag{A.69}$$

are bounds on the bias and variance respectively, and

$$K^* = \min\{k \geq 0 : r_k(\Sigma_X)/n \geq b\}, \tag{A.70}$$

where $b > 1$ is an absolute constant.

We now compare these two terms to the corresponding terms in our bound in Theorem 13.

Comparison of Variance Terms

We first compare the variance term $\boldsymbol{V}$ to corresponding variance term in our Theorem 13, display (1.27). Note that as long as the SNR

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

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 32. *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 32 and for n large enough,

$$\boldsymbol{V} := \sigma^2 \log(n)\left(\frac{n}{R_K(\Sigma_X)} + \frac{K}{n}\right).$$

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

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

Thus,

$$V \geq \sigma^2 \log(n)\left(\frac{n}{p} + \frac{K}{n}\right). \tag{A.71}$$

When $\kappa(\Sigma_E) < c_1$, $p \lesssim \mathrm{r_e}(\Sigma_E) \leq p$, and so the variance term in the bound of our Theorem 13 is

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

Thus, comparing with (A.71), we see that under the stated conditions our variance bound is the same as that of [13], up to absolute constants.

Proof of Lemma 32. We will prove that

$$\frac{r_\ell(\Sigma_X)}{n} \leq \frac{K}{n}(1 + \xi^{-1}) + \frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}, \qquad \text{for } 0 \leq \ell \leq K-1 \tag{A.72}$$

and that

$$\frac{r_K(\Sigma_X)}{n} \geq \frac{\mathrm{r_e}(\Sigma_E)}{n} - \frac{K}{n}. \tag{A.73}$$

Together with the definition of K^* in (A.70), these two bounds imply Lemma 32.

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

$$\begin{aligned}
\sum_{i=\ell+1}^{p} \lambda_i(\Sigma_X) &= \mathrm{tr}(\Sigma_X) - \sum_{i=1}^{\ell} \lambda_i(\Sigma_X) \\
&= \mathrm{tr}(\Sigma_E) + \mathrm{tr}(A\Sigma_Z A^\top) - \sum_{i=1}^{\ell} \lambda_i(\Sigma_X) \\
&= \mathrm{tr}(\Sigma_E) + \sum_{i=\ell+1}^{K} \lambda_i(A\Sigma_Z A^\top) + \sum_{i=1}^{\ell} (\lambda_i(A\Sigma_Z A^\top) - \lambda_i(\Sigma_X)),
\end{aligned} \tag{A.74}$$

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 (A.72): By Weyl's inequality,

$$|\lambda_i(A\Sigma_Z A^\top) - \lambda_i(\Sigma_X)| \leq \|\Sigma_E\|, \tag{A.75}$$

so by (A.74),

$$\sum_{i=\ell+1}^{p} \lambda_i(\Sigma_X) \le \operatorname{tr}(\Sigma_E) + (K-\ell)\lambda_{\ell+1}(A\Sigma_Z A^\top) + \ell\|\Sigma_E\|$$
$$\le \operatorname{tr}(\Sigma_E) + K\lambda_{\ell+1}(A\Sigma_Z A^\top) + K\|\Sigma_E\|. \tag{A.76}$$

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

$$\lambda_{\ell+1}(\Sigma_X) \ge \lambda_{\ell+1}(A\Sigma_Z A^\top). \tag{A.77}$$

Combining (A.76) and (A.77), 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 (A.72).

Proof of (A.73): Equation (A.74) 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 (A.75),

$$\sum_{i=K+1}^{p} \lambda_i(\Sigma_X) \ge \operatorname{tr}(\Sigma_E) - K\|\Sigma_E\|. \tag{A.78}$$

Since

$$\begin{aligned}\lambda_{K+1}(\Sigma_X) &= \lambda_{K+1}(\Sigma_X) - \lambda_{K+1}(A\Sigma_Z A^\top) \qquad (\text{since } \lambda_{K+1}(A\Sigma_Z A^\top) = 0) \\ &\leq \|\Sigma_E\| \qquad (\text{Weyl's inequality}). \end{aligned} \tag{A.79}$$

Combining (A.78) and (A.79), 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 (A.73).

■

Comparison of Bias Terms

A more interesting comparison arises between the bias term B and the corresponding bias term in Theorem 13, display (1.27). 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 33. *Suppose $\xi := \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_E\| > 1$ and A, Σ_Z, Σ_E are all full rank. Then*

$$B \geq \left(\frac{\xi - 1}{\xi + 1}\right) \cdot \frac{1}{\kappa(\Sigma_E)} \|\beta\|_{\Sigma_Z}^2 \max\left(\sqrt{\frac{r_0(\Sigma_X)}{n}}, \frac{r_0(\Sigma_X)}{n}\right), \tag{A.80}$$

where

$$\frac{r_0(\Sigma_X)}{n} \geq \frac{1}{2}\frac{r_0(A\Sigma_Z A^\top)}{n} + \frac{1}{2\kappa(A\Sigma_Z A^\top)}\frac{1}{\xi}\frac{\mathrm{r_e}(\Sigma_E)}{n}. \tag{A.81}$$

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,

$$B \gtrsim \|\beta\|_{\Sigma_Z}^2 \max\left(\sqrt{\frac{1}{\xi}\frac{p}{n}}, \frac{1}{\xi}\frac{p}{n}\right). \tag{A.82}$$

Compared to our bias bound $\|\beta\|_{\Sigma_Z}^2 p/(n \cdot \xi)$ in Theorem 13, 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 (A.82). 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 13.

Proof of Lemma 33. Using that A, Σ_Z, Σ_E are all full rank, by (A.28) 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 (A.80).

To prove (A.81), 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 $\|\Sigma_X\| \leq \|A\Sigma_Z A^\top\| + \|\Sigma_E\| \leq 2\|A\Sigma_Z A^\top\|$, where we use that $\|\Sigma_E\| \leq \|A\Sigma_Z A^\top\|$ by the assumption $\xi > 1$, we find

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

which proves (A.81). ■

A.4 Supplementary Results

A.4.1 Closed Form Solutions of Min-Norm Estimator and Minimizer of $R(\alpha)$

Lemma 34. *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 (A.5), $\Sigma_X \alpha^* = \Sigma_{XY}$, so the claim is proved.

■

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

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

We then have the following result.

Lemma 35. $\widehat{\alpha} = \boldsymbol{X}^+ \mathbf{y}$.

Proof. We establish the proof in two steps.

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

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

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

$$\boldsymbol{X}^\top \boldsymbol{X}u = \boldsymbol{X}^\top \mathbf{y}. \tag{A.83}$$

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

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

and thus $\boldsymbol{X}^+\mathbf{y}$ is a minimizer of $\|\boldsymbol{X}u - \mathbf{y}\|$. The set of vectors u satisfying $\boldsymbol{X}^\top \boldsymbol{X}u = \boldsymbol{X}^\top \mathbf{y}$ 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 $\|\boldsymbol{X}u - \mathbf{y}\|$, it must satisfy A.83, i.e.

$$\boldsymbol{X}^\top \boldsymbol{X}\widehat{\alpha} = \boldsymbol{X}^\top y. \tag{A.84}$$

We can write

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

and using $\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{X}$ as well as the fact that $\boldsymbol{X}^+\boldsymbol{X}$ is symmetric (see Appendix A.5), a quick calculation gives

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

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

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

where we used $\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{X}$ in the first step and A.84 in the second step. Thus $\boldsymbol{X}^+\boldsymbol{X}\widehat{\alpha}$ is a minimizer of $\|\cdot\|$ among minimizers of $\|\boldsymbol{X}u - \mathbf{y}\|$. Since by Step 1 above $\widehat{\alpha}$ is the unique such minimizer, $\boldsymbol{X}^+\boldsymbol{X}\widehat{\alpha} = \widehat{\alpha}$. Thus,

$$\begin{aligned}
\widehat{\alpha} &= \boldsymbol{X}^+\boldsymbol{X}\widehat{\alpha} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\boldsymbol{X}\widehat{\alpha} && \text{(since } \boldsymbol{X}^+ = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\text{)} \\
&= (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\mathbf{y} && \text{(by A.84)} \\
&= \boldsymbol{X}^+\mathbf{y}. && \text{(since } \boldsymbol{X}^+ = (\boldsymbol{X}^\top\boldsymbol{X})^+\boldsymbol{X}^\top\text{)}
\end{aligned}$$

■

A.4.2 Proof that (1.5) is a Special Case of (1.21) in the Gaussian Case

Lemma 36. *Suppose that (X, y) follows model (1.5) with mean zero and is furthermore jointly Gaussian. Then model (1.21) 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 (1.5).*

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 (A.5) 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 (1.21) holds as claimed. ■

A.4.3 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 37. *Under model 1.5, 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 $1 - c/n$,*

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

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 13 found in section A.3.2 above,

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

where

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

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

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

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 $\boldsymbol{Z}$ is full rank, $\boldsymbol{Z}^+\boldsymbol{Z} = I_K$ and so

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

Thus,

$$\begin{aligned} B_2 &= \|(A^\top \boldsymbol{X}^+\boldsymbol{Z} - I_K)\beta\|^2 \\ &= \|(\boldsymbol{Z}^+\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{Z} - I_K)\beta - \boldsymbol{Z}^+\mathcal{E}\boldsymbol{X}^+\boldsymbol{Z}\beta\|_{\Sigma_Z}^2 \\ &\leq 2\|(\boldsymbol{Z}^+\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{Z} - I_K)\beta\|_{\Sigma_Z}^2 + 2\|\boldsymbol{Z}^+\mathcal{E}\boldsymbol{X}^+\boldsymbol{Z}\beta\|_{\Sigma_Z}^2. \end{aligned} \tag{A.85}$$

Note that since $p \geq K$, by Assumption 2, $\operatorname{rank}(A) = K$ so by Lemma 39 of Appendix A.5,

$$A^\top A^{+\top} = I_K. \tag{A.86}$$

We thus have

$$\begin{aligned} \|(\boldsymbol{Z}^+\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{Z} - I_K)\beta\|_{\Sigma_Z}^2 &= \|(\boldsymbol{Z}^+\boldsymbol{X}\boldsymbol{X}^+\boldsymbol{Z} - \boldsymbol{Z}^+\boldsymbol{Z})\beta\|_{\Sigma_Z}^2 \\ &= \|\tilde{\boldsymbol{Z}}^+(\boldsymbol{X}\boldsymbol{X}^+ - I_p)\boldsymbol{Z}\beta\|^2 \\ &\leq \frac{\|(\boldsymbol{X}\boldsymbol{X}^+ - I_p)\boldsymbol{Z}\beta\|^2}{\sigma_K^2(\tilde{\boldsymbol{Z}})} \\ &\lesssim \frac{1}{n}\|(\boldsymbol{X}\boldsymbol{X}^+ - I_p)\boldsymbol{Z}\beta\|^2 && \text{(on } \mathcal{B}\text{)} \\ &= \frac{1}{n}\|(\boldsymbol{X}\boldsymbol{X}^+ - I_p)\boldsymbol{Z}A^\top A^{+\top}\beta\|^2 && \text{(by (A.86))} \\ &= \frac{1}{n}\|(\boldsymbol{X}\boldsymbol{X}^+ - I_p)(\boldsymbol{X} - \mathcal{E})A^{+\top}\beta\|^2 && \text{(since } \boldsymbol{X} = \boldsymbol{Z}A^\top + \mathcal{E}\text{)} \\ &= \frac{1}{n}\|(\boldsymbol{X}\boldsymbol{X}^+ - I_p)\mathcal{E}A^{+\top}\beta\|^2 && \text{(since } \boldsymbol{X}\boldsymbol{X}^+\boldsymbol{X} = \boldsymbol{X}\text{)} \\ &\leq \frac{1}{n}\|\boldsymbol{X}\boldsymbol{X}^+ - I_p\| \cdot \|\mathcal{E}A^{+\top}\beta\|^2 \\ &\leq \frac{1}{n}\|\mathcal{E}A^{+\top}\beta\|^2 \\ &\lesssim \frac{n\|\Sigma_E\|}{n}\frac{\|\beta\|_{\Sigma_Z}^2}{\lambda_K(A\Sigma_Z A^\top)} && \text{(on } \mathcal{B} \text{ and by (A.47))} \\ &= \frac{\|\beta\|_{\Sigma_Z}^2}{\xi}, \end{aligned} \tag{A.87}$$

where in the penultimate step we used

$$\|A^{+\top}\beta\|^2 \leq \frac{\|\beta\|^2_{\Sigma_Z}}{\lambda_K(A\Sigma_Z A^\top)} \tag{A.88}$$

from (A.47). We can bound the second term in A.85 as follows:

$$\begin{aligned}
\|\boldsymbol{Z}^+\mathcal{E}\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2_{\Sigma_Z} &= \|\tilde{\boldsymbol{Z}}^+\mathcal{E}\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2 \\
&\leq \frac{\|\mathcal{E}\|^2}{\sigma_K^2(\tilde{\boldsymbol{Z}})}\|\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2 \\
&\lesssim \|\Sigma_E\| \cdot \|\boldsymbol{X}^+\boldsymbol{Z}\beta\|^2 && \text{(on } \mathcal{B}\text{)} \\
&= \|\Sigma_E\| \cdot \|\boldsymbol{X}^+\boldsymbol{Z}A^\top A^{+\top}\beta\|^2 && \text{(since } A^\top A^{+\top} = I_K\text{)} \\
&= \|\Sigma_E\| \cdot \|\boldsymbol{X}^+(\boldsymbol{X} - \mathcal{E})A^{+\top}\beta\|^2 && \text{(since } \boldsymbol{X} = \boldsymbol{Z}A^\top + \mathcal{E}\text{)} \\
&\leq 2\|\Sigma_E\| \cdot \|\boldsymbol{X}^+\boldsymbol{X}A^{+\top}\beta\|^2 + 2\|\Sigma_E\| \cdot \|\boldsymbol{X}^+\mathcal{E}A^{+\top}\beta\|^2 \\
&\lesssim \|\Sigma_E\|\|A^{+\top}\beta\|^2 + \|\Sigma_E\|\frac{\|\mathcal{E}\|}{\sigma_p^2(\boldsymbol{X})}\|A^{+\top}\beta\|^2 && \text{(since } \|\boldsymbol{X}^+\boldsymbol{X}\| \leq 1\text{)} \\
&\lesssim \|\Sigma_E\| \cdot \kappa(\Sigma_E)\|A^{+\top}\beta\|^2 \\
&\leq \kappa(\Sigma_E)\frac{\|\beta\|^2_{\Sigma_Z}}{\xi}. && \text{(by (A.88))}
\end{aligned}$$

Using this and (A.87) in (A.85), and using the fact that $\kappa(\Sigma_E) > 1$, we find that on the event $\mathcal{B}$,

$$B_2 \lesssim \kappa(\Sigma_E)\frac{\|\beta\|^2_{\Sigma_Z}}{\xi}. \tag{A.89}$$

Bounding the variance component: We have

$$\begin{aligned}
V_1 + V_2 &= \boldsymbol{\varepsilon}^\top\boldsymbol{X}^{+\top}\Sigma_X\boldsymbol{X}^+\boldsymbol{\varepsilon} \\
&= \sigma^2\tilde{\boldsymbol{\varepsilon}}^\top\boldsymbol{X}^{+\top}\Sigma_X\boldsymbol{X}^+\tilde{\boldsymbol{\varepsilon}} && \text{(by Assumption 3)} \\
&\lesssim \sigma^2\log(n)\mathrm{tr}(\boldsymbol{X}^{+\top}\Sigma_X\boldsymbol{X}^+) && \text{(on } \mathcal{B}_2\text{)} \\
&\leq \sigma^2\log(n) \cdot p\|\boldsymbol{X}^{+\top}\Sigma_X\boldsymbol{X}^+\| && \text{(since rank}(\boldsymbol{X}^+) = p\text{)} \\
&= \sigma^2\log(n) \cdot p\|\Sigma_X^{1/2}\boldsymbol{X}^+\|^2.
\end{aligned} \tag{A.90}$$

From Assumption 1, $\boldsymbol{X} = \tilde{\mathbf{X}}\Sigma_X^{1/2}$, and from Lemma 39 of Appendix A.5 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 (A.90), we find

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

Proof that $\mathbb{P}(\mathcal{B}) \geq 1 - c/n$*:* The bounds $\mathbb{P}(\mathcal{B}_1) \geq 1 - c/n$ and $\mathbb{P}(\mathcal{B}_2) \geq 1 - e^{-cn}$ follow respectively from Theorem 4.6.1 of [91] and Lemma 30 in Appendix A.3.1 above, by similar reasoning as in the proof of Theorem 13, for example. ■

A.4.4 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 $1 \leq a \leq K$. We recall the definition

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

Lemma 38. $\xi \geq \min_a |I_a| \cdot \lambda_K(\Sigma_Z)/\|\Sigma_E\|$.

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. ■

A.5 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 [83], 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:

$$BB^+B = B \tag{A.91}$$

$$B^+BB^+ = B^+ \tag{A.92}$$

$$BB^+ \text{ is symmetric} \tag{A.93}$$

$$B^+B \text{ is symmetric} \tag{A.94}$$

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

Lemma 39. *For any $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{m\times d}$,*

$$(BC)^+ = (B^+BC)^+(BCC^+)^+. \tag{A.95}$$

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

$$B^\top BB^+ = B^\top \tag{A.96}$$

$$B^\top (BB^\top)^+ = B^+ \tag{A.97}$$

$$(B^\top B)^+ B^\top = B^+ \tag{A.98}$$

$$B^+B = I_m \text{ if } r = m \tag{A.99}$$

$$BB^+ = I_n \text{ if } r = n \tag{A.100}$$

$$\|B^+\| = 1/\sigma_r(B) \tag{A.101}$$

$$rank(B^+) = rank(B) = r. \tag{A.102}$$

APPENDIX B

APPENDIX OF CHAPTER 2

B.1 Organization of Appendices

We provide section-by-section proofs for the main results in Appendices B.2.1—B.2.4. Auxiliary lemmas are collected in Appendix B.3. Appendix B.4 contains the procedure of estimating A under the Essential Regression framework while comparison with more existing literature on factor models is stated in Appendix B.5.

B.2 Main proofs

We start by giving an elementary lemma that proves $Y^*_{\widehat{B}} = Y^*_{P_{\widehat{B}}}$ for any $\widehat{B} \in \mathbb{R}^{p\times q}$. Recall that, for any matrix M, M^+ denotes its Moore-Penrose inverse and P_M denotes the projection onto the column space of M.

Lemma 40. *Let $\widehat{B} \in \mathbb{R}^{p\times q}$ be any matrix. Then*

$$\widehat{B}(\boldsymbol{X}\widehat{B})^+ = P_{\widehat{B}}(\boldsymbol{X}P_{\widehat{B}})^+.$$

Proof. Write the SVD of $\widehat{B}$ as $\widehat{B} = UDV^\top$ where $U \in \mathbb{R}^{p\times r_0}$ and $V \in \mathbb{R}^{q\times r_0}$ are

orthonormal matrices with $r_0 = \text{rank}(\widehat{B})$. We then have

$$\begin{aligned}
\widehat{B}(\boldsymbol{X}\widehat{B})^+ &= \widehat{B}\left(\widehat{B}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{B}\right)^+ \widehat{B}^\top \boldsymbol{X}^\top \\
&= UDV^\top \left(VDU^\top \boldsymbol{X}^\top \boldsymbol{X} UDV^\top\right)^+ VDU^\top \boldsymbol{X}^\top \\
&\overset{(i)}{=} U(U^\top \boldsymbol{X}^\top \boldsymbol{X}U)^+ U^\top \boldsymbol{X}^\top \\
&\overset{(ii)}{=} UU^\top (UU^\top \boldsymbol{X}^\top \boldsymbol{X} UU^\top)^+ UU^\top \boldsymbol{X}^\top.
\end{aligned}$$

The result then follows by noting that $P_{\widehat{B}} = UU^\top$. Step (i) uses the fact that

$$\left(VDU^\top \boldsymbol{X}^\top \boldsymbol{X} UDV^\top\right)^+ = VD^{-1}\left(U^\top \boldsymbol{X}^\top \boldsymbol{X}U\right)^+ D^{-1}V^\top$$

which can be verified by the definition of Moore-Penrose inverse. Indeed, let $M = U^\top \boldsymbol{X}^\top \boldsymbol{X}U$, $N = VDMDV^\top$ and $\widetilde{N} = VD^{-1}M^+D^{-1}V^\top$. We need to verify

$$N\widetilde{N}N = N, \qquad \widetilde{N}N\widetilde{N} = \widetilde{N}.$$

Straightforwardly,

$$N\widetilde{N}N = VDMM^+MDV^\top = VDMDV^\top = N$$

and similar arguments hold for $\widetilde{N}N\widetilde{N} = \widetilde{N}$. Step (ii) uses step (i) with $D = \boldsymbol{I}_{r_0}$ and $V = U$ ■

B.2.1 Proofs for Section 2.2

Proof of Lemma 16

Let $\Sigma_X = \text{Cov}(X)$, $\Sigma_{XY} = \text{Cov}(X, Y)$. Since Σ_W is invertible, $\lambda_p(\Sigma_X) = \lambda_p(A\Sigma_Z A^\top + \Sigma_W) \geq \lambda_p(\Sigma_W) > 0$ so Σ_X is invertible. Thus, letting $\alpha^* = \Sigma_X^{-1}\Sigma_{XY}$,

$$\mathbb{R}^* - \sigma^2 = \mathbb{E}[(X^\top \alpha^* - Z^\top \beta)^2]. \tag{B.1}$$

Using this expression, and the factor model structure $X = AZ + W$, $Y = Z^\top\beta + \varepsilon$, the proof of Lemma 4 in [33] uses the Woodbury matrix identity to simplify (B.1), arriving at

$$\mathbb{R}^* - \sigma^2 = \beta^\top(\Sigma_Z^{-1} + A^\top\Sigma_W^{-1}A)^{-1}\beta.$$

Letting $H = \Sigma_Z^{1/2}A^\top\Sigma_W^{-1}A\Sigma_Z^{1/2}$, we then have

$$\begin{aligned}\mathbb{R}^* - \sigma^2 &= \beta^\top\Sigma_Z^{1/2}(\boldsymbol{I}_K + H)^{-1}\Sigma_Z^{1/2}\beta \\ &= \beta^\top\Sigma_Z^{1/2}H^{-1/2}(\boldsymbol{I}_K + H^{-1})^{-1}H^{-1/2}\Sigma_Z^{1/2}\beta.\end{aligned}$$

To obtain the upper bound on $\mathbb{R}^*$ we use

$$\mathbb{R}^* - \sigma^2 = \beta^\top\Sigma_Z^{1/2}H^{-1/2}(\boldsymbol{I}_K + H^{-1})^{-1}H^{-1/2}\Sigma_Z^{1/2}\beta \le \frac{\beta^\top\Sigma_Z^{1/2}H^{-1}\Sigma_Z^{1/2}\beta}{1 + \lambda_K(H^{-1})} \le \beta^\top(A^\top\Sigma_W^{-1}A)^{-1}\beta,$$

where we used $\Sigma_Z^{1/2}H^{-1}\Sigma_Z^{1/2} = (A^\top\Sigma_X^{-1}A)^{-1}$ in the last step.

To find the lower bound we first observe that

$$\mathbb{R}^* - \sigma^2 = \beta^\top\Sigma_Z^{1/2}H^{-1/2}(\boldsymbol{I}_K + H^{-1})^{-1}H^{-1/2}\Sigma_Z^{1/2}\beta \ge \frac{\beta^\top\Sigma_Z^{1/2}H^{-1}\Sigma_Z^{1/2}\beta}{1 + \|H^{-1}\|_{\mathrm{op}}} = \frac{\beta^\top(A^\top\Sigma_X^{-1}A)^{-1}\beta}{1 + \lambda_K^{-1}(H)}.$$

Furthermore,

$$\lambda_K(H) = \lambda_K(\Sigma_Z^{1/2}A^\top\Sigma_W^{-1}A\Sigma_Z^{1/2}) \ge \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_W\|_{\mathrm{op}} = \xi,$$

so using this in the previous display,

$$\mathbb{R}^* - \sigma^2 \ge \frac{\beta^\top(A^\top\Sigma_X^{-1}A)^{-1}\beta}{1 + \xi^{-1}} = \frac{\xi}{1 + \xi} \cdot \beta^\top(A^\top\Sigma_X^{-1}A)^{-1}\beta,$$

as claimed. ■

Proof of Theorem 17

Define $\widehat{\alpha}_{\widehat{B}} = \widehat{B}\left(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B}\right)^+\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{Y}$ and recall that $\widehat{Y}^*_{\widehat{B}} = X_*^\top\widehat{\alpha}_{\widehat{B}}$ from (2.3). Pick any θ with $K \le (Cn/\log n) \wedge p$ such that (X, Y) follows FRM(θ) where $C = C(\gamma_z)$ is some

positive constant. By $X_* = AZ_* + W_*$ and $Y_* = Z_*^\top \beta + \varepsilon_*$, and the independence of Z_*, ε_*, and W_*, one has

$$\begin{aligned}
\mathbb{R}(\widehat{B}) - \sigma^2 &= \mathbb{E}_{(Z_*, W_*)}\left[\left(\widehat{Y}^*_{\widehat{B}} - Z_*^\top \beta\right)^2\right] \\
&= \mathbb{E}_{Z_*}\left[\left(Z_*^\top A^\top \widehat{\alpha}_{\widehat{B}} - Z_*^\top \beta\right)^2\right] + \mathbb{E}_{W_*}\left[\left(W_*^\top \widehat{\alpha}_{\widehat{B}}\right)^2\right] \qquad \text{(B.2)} \\
&= \left\|\Sigma_Z^{1/2}\left(A^\top \widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2 + \left\|\Sigma_W^{1/2} \widehat{\alpha}_{\widehat{B}}\right\|^2 \\
&\le \left\|\Sigma_Z^{1/2}\left(A^\top \widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2 + \|\Sigma_W\|_{\mathrm{op}} \left\|\widehat{\alpha}_{\widehat{B}}\right\|^2. \qquad \text{(B.3)}
\end{aligned}$$

We define an event $\mathcal{E}^*$ in (B.4) below, on which we bound the risk. Invoking Lemmas 42, 43 and using $\beta^\top A^+ \Sigma_W A^{+\top} \beta \le \beta^\top (A^\top A)^{-1} \beta \|\Sigma_W\|_{\mathrm{op}}$, we find that the stated bound holds on the event $\mathcal{E}^*$. Then, by Lemma 41, $\mathbb{P}(\mathcal{E}^*) \ge 1 - cn^{-1}$, which completes the proof. ■

We state and prove three lemmas which are used in the proof of Theorem 17. Recall that

$$\widehat{r} = \mathrm{rank}(\boldsymbol{X} P_{\widehat{B}}), \qquad \widehat{\psi} = \frac{1}{n}\sigma_1^2\left(\boldsymbol{X} P^\perp_{\widehat{B}}\right), \qquad \widehat{\eta} = \frac{1}{n}\sigma^2_{\widehat{r}}\left(\boldsymbol{X} P_{\widehat{B}}\right).$$

Lemma 41. *For any θ with $K \le (Cn/\log n) \wedge p$ and some positive constant $C = C(\gamma_z)$ such that (X, Y) follows FRM(θ), we have $\mathbb{P}(\mathcal{E}^*) \ge 1 - cn^{-1}$ for some absolute constant $c > 0$, where we define the event*

$$\mathcal{E}^* := \mathcal{E}_{\boldsymbol{Z}} \cap \mathcal{E}_{\boldsymbol{W}} \cap \mathcal{E}'_{\boldsymbol{W}} \cap \mathcal{E}_M \cap \mathcal{E}_{M'} \cap \mathcal{E}_{\boldsymbol{Z}\beta}. \qquad \text{(B.4)}$$

Here, for some constants $c(\gamma_z)$ and $c'(\gamma_w)$ depending on γ_z and γ_w, respectively,

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{Z}} &:= \left\{ \lambda_K \left(\Omega^{1/2} \frac{1}{n} \boldsymbol{Z}^\top \boldsymbol{Z} \Omega^{1/2} \right) \geq c(\gamma_z) \right\}, \\
\mathcal{E}_{\boldsymbol{Z}\beta} &:= \left\{ \frac{1}{n} \left\| P^{\perp}_{\boldsymbol{X}\widehat{B}} \boldsymbol{Z}\beta \right\|^2 \leq 8\gamma_w^2 \beta^\top A^+ \Sigma_W A^{+\top} \beta + 2\widehat{\psi} \beta^\top (A^\top A)^{-1} \beta \right\}, \\
\mathcal{E}_{\boldsymbol{W}} &:= \left\{ \frac{1}{n} \left\| \boldsymbol{W}^\top \boldsymbol{W} \right\|_{op} \leq \delta_W \right\}, \\
\mathcal{E}'_{\boldsymbol{W}} &:= \left\{ \frac{1}{n} \left\| \boldsymbol{W} A^{+\top} \beta \right\|^2 \leq 4\gamma_w^2 \beta^\top A^+ \Sigma_W A^{+\top} \beta \right\}, \\
\mathcal{E}_M &:= \left\{ \boldsymbol{\varepsilon}^\top M \boldsymbol{\varepsilon} \leq 2\gamma_\varepsilon^2 \sigma^2 \left[2\|M\|_{op} \log n + tr(M) \right] \right\}, \\
\mathcal{E}_{M'} &:= \left\{ \boldsymbol{\varepsilon}^\top M' \boldsymbol{\varepsilon} \leq 2\gamma_\varepsilon^2 \sigma^2 \left[2\|M'\|_{op} \log n + tr(M') \right] \right\},
\end{aligned}$$

with $\Omega := \Sigma_Z^{-1}$, δ_W defined in (2.12), and

$$\begin{aligned}
M &:= (\boldsymbol{X}\widehat{B})^{+\top} \widehat{B}^\top \widehat{B} (\boldsymbol{X}\widehat{B})^+, \\
M' &:= (\boldsymbol{X}\widehat{B})^{+\top} \widehat{B}^\top A \Sigma_Z A^\top \widehat{B} (\boldsymbol{X}\widehat{B})^+.
\end{aligned}$$

Proof. By an application of Theorem 5.39 of [90] and $K \log n \leq C(\gamma_z) n$, we find $\mathbb{P}\{\mathcal{E}_{\boldsymbol{Z}}^c\} \lesssim n^{-c'K}$. From Lemma 47 with $\boldsymbol{G} = \boldsymbol{W}\Sigma_W^{-1/2}$, $H = \Sigma_W$, and $\gamma = \gamma_w$, we find $\mathbb{P}\{\mathcal{E}_{\boldsymbol{W}}^c\} \leq e^{-n}$.

We note that $\boldsymbol{W}A^{+\top}\beta$ has independent $\gamma_w \sqrt{\beta^\top A^+ \Sigma_W A^{+\top} \beta}$ sub-Gaussian entries, so $\boldsymbol{W}A^{+\top}\beta$ is a $\gamma_w \sqrt{\beta^\top A^+ \Sigma_W A^{+\top} \beta}$ sub-Gaussian random vector. Applying Lemma 46 with $\xi = \boldsymbol{W}A^{+\top}\beta$, $H = \boldsymbol{I}_n$, $\gamma_\xi^2 = \gamma_w^2 \beta^\top A^+ \Sigma_W A^{+\top} \beta$ and choosing $t = \log n$ yield

$$\mathbb{P}\{(\mathcal{E}'_{\boldsymbol{W}})^c\} = \mathbb{P}\left\{ \frac{1}{n} \left\| \boldsymbol{W} A^{+\top} \beta \right\|^2 > 4\gamma_w^2 \beta^\top A^+ \Sigma_W A^{+\top} \beta \right\} \leq n^{-1}. \tag{B.5}$$

We prove $\mathcal{E}'_{\boldsymbol{W}} \cap \mathcal{E}_{\boldsymbol{Z}\beta} = \mathcal{E}_{\boldsymbol{W}'}$ in Lemma 44. By the independence of $\boldsymbol{\varepsilon}$ and both $\boldsymbol{X}$ and $\widehat{B}$, the matrix M is independent of $\boldsymbol{\varepsilon}$. Thus, by an application of Lemma 46 with $\xi = \boldsymbol{\varepsilon}$, $H = M$, $\gamma_\xi = \sigma\gamma_\varepsilon$ and $t = \log n$ gives $\mathbb{P}\{\mathcal{E}_M^c | M\} \leq n^{-1}$. Taking the expectation over M then gives $\mathbb{P}\{\mathcal{E}_M^c\} \leq n^{-1}$. The same argument with $H = M'$ gives $\mathbb{P}\{\mathcal{E}_{M'}^c\} \leq n^{-1}$.

Combining results, we find

$$\mathbb{P}\{\mathcal{E}^{*c}\} \le \mathbb{P}\{\mathcal{E}_{\boldsymbol{Z}}^c\} + \mathbb{P}\{\mathcal{E}_{\boldsymbol{W}}^c\} + \mathbb{P}\{(\mathcal{E}_{\boldsymbol{W}}')^c\} + \mathbb{P}\{\mathcal{E}_M^c\} + \mathbb{P}\{\mathcal{E}_{M'}^c\} \lesssim n^{-1}.$$

■

Lemma 42. *Under conditions of Theorem 17, on the event $\mathcal{E}^*$ defined in (B.4),*

$$\|\widehat{\alpha}_{\widehat{B}}\|^2 \lesssim_\theta \frac{(\widehat{r} + \log n)\sigma^2}{n\widehat{\eta}} + \beta^\top (A^\top A)^{-1}\beta + \widehat{\eta}^{-1}\left(\widehat{\psi}\beta^\top (A^\top A)^{-1}\beta + \beta^\top A^+ \Sigma_W A^{+\top}\beta\right). \quad \text{(B.6)}$$

Proof. Starting with the identity

$$\widehat{\alpha}_{\widehat{B}} = \widehat{B}(\boldsymbol{X}\widehat{B})^+ \boldsymbol{Y} = \widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{Z}\beta + \boldsymbol{\varepsilon}), \quad \text{(B.7)}$$

with $(\boldsymbol{X}\widehat{B})^+ := (\widehat{B}\boldsymbol{X}^\top \boldsymbol{X}\widehat{B})^+ \widehat{B}^\top \boldsymbol{X}^\top$, we have

$$\|\widehat{\alpha}_{\widehat{B}}\|^2 \le 2\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 + 2\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2.$$

To bound the first term, notice that

$$\begin{aligned}
\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 &= \boldsymbol{\varepsilon}^\top (\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}^\top M \boldsymbol{\varepsilon} \\
&\le 2\gamma_\varepsilon^2\sigma^2\Big[2\|M\|_{\mathrm{op}}\log n + \mathrm{tr}(M)\Big],
\end{aligned}$$

where the last step holds on $\mathcal{E}^*$ (in particular, on $\mathcal{E}_M \subset \mathcal{E}^*$). Observe that, on $\mathcal{E}^*$,

$$\begin{aligned}
\mathrm{tr}(M) &= \mathrm{tr}\left((\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\right) \\
&\le \mathrm{rank}(\boldsymbol{X}\widehat{B}) \cdot \|M\|_{\mathrm{op}} \\
&= \widehat{r}\|M\|_{\mathrm{op}}.
\end{aligned}$$

Write the SVD of $\widehat{B}$ as $\widehat{B} = UDV^\top$ where $U \in \mathbb{R}^{p\times r_0}$ and $V \in \mathbb{R}^{q\times r_0}$ are orthogonal matrices with $r_0 = \mathrm{rank}(\widehat{B})$. Recalling that $(\boldsymbol{X}\widehat{B})^+ = (\widehat{B}^\top \boldsymbol{X}^\top \boldsymbol{X}\widehat{B})^+ \widehat{B}\boldsymbol{X}^\top$, the following

holds, on the event $\mathcal{E}^*$,

$$\begin{aligned}
\|M\|_{\mathrm{op}} &= \left\|(\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}} \\
&\overset{(i)}{=} \left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\right\|_{\mathrm{op}} \\
&= \left\|\widehat{B}(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B})^+\widehat{B}\boldsymbol{X}^\top\boldsymbol{X}\widehat{B}(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B})^+\widehat{B}^\top\right\|_{\mathrm{op}} \\
&= \left\|\widehat{B}\left(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B}\right)^+\widehat{B}^\top\right\|_{\mathrm{op}} \\
&= \left\|U\left(U^\top\boldsymbol{X}^\top\boldsymbol{X}U\right)^+U^\top\right\|_{\mathrm{op}} \\
&\overset{(ii)}{\leq} \sigma_{\widehat{r}}^{-2}(\boldsymbol{X}U) \\
&\overset{(iii)}{=} (n\widehat{\eta})^{-1}
\end{aligned} \tag{B.8}$$

where we used $\|FF^\top\|_{\mathrm{op}} = \|F^\top F\|_{\mathrm{op}}$ for any matrix F in (i), $\mathrm{rank}(\boldsymbol{X}U) = \mathrm{rank}(\boldsymbol{X}P_{\widehat{B}}) = \widehat{r}$ in (ii) and

$$\sigma_{\widehat{r}}^2(\boldsymbol{X}U) = \lambda_{\widehat{r}}(\boldsymbol{X}UU^\top\boldsymbol{X}) = \lambda_{\widehat{r}}(\boldsymbol{X}P_{\widehat{B}}^2\boldsymbol{X}) = \sigma_{\widehat{r}}(\boldsymbol{X}P_{\widehat{B}})$$

in (iii). This concludes, on the event $\mathcal{E}^*$,

$$\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 \leq \frac{2\gamma_\varepsilon^2\sigma^2}{n\widehat{\eta}}(\widehat{r} + 2\log n). \tag{B.9}$$

On the other hand, by $A^\top A^{+\top} = \boldsymbol{I}_K$ and $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$, observe that

$$\begin{aligned}
\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} &= \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}A^\top A^{+\top} \\
&= \widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{X} - \boldsymbol{W})A^{+\top} \\
&= \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}A^{+\top} + \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}^\perp A^{+\top} - \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{W}A^{+\top}.
\end{aligned} \tag{B.10}$$

By $P_{\widehat{B}} = \widehat{B}\widehat{B}^+$ and the inequality $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$,

$$\begin{aligned}
\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2 &\leq 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+A^{+\top}\beta\right\|^2 + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}^\perp A^{+\top}\beta\right\|^2 \\
&\quad + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{W}A^{+\top}\beta\right\|^2 \\
&\leq 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+\right\|_{\mathrm{op}}^2\left\|A^{+\top}\beta\right\|^2 + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\left\|\boldsymbol{X}P_{\widehat{B}}^\perp\right\|_{\mathrm{op}}^2\left\|A^{+\top}\beta\right\|^2 \\
&\quad + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2.
\end{aligned} \tag{B.11}$$

Recalling $\widehat{B} = UDV^\top$, on the event $\mathcal{E}^*$, the following observation

$$\begin{aligned}\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+\right\|_{\mathrm{op}} &= \left\|U\left(U^\top\boldsymbol{X}^\top\boldsymbol{X}U\right)^+U^\top\boldsymbol{X}^\top\boldsymbol{X}UU^\top\right\|_{\mathrm{op}}\\ &\le \left\|\left(U^\top\boldsymbol{X}^\top\boldsymbol{X}U\right)^+U^\top\boldsymbol{X}^\top\boldsymbol{X}U\right\|_{\mathrm{op}} \le 1,\end{aligned}$$

together with (B.8), concludes

$$\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2 \le 3\beta^\top(A^\top A)^{-1}\beta + 3\widehat{\eta}^{-1}\left(\widehat{\psi}\beta^\top(A^\top A)^{-1}\beta + 4\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta\right). \tag{B.12}$$

Collecting (B.9)—(B.12) concludes the proof. ■

Lemma 43. *Under conditions of Theorem 17, on the event $\mathcal{E}^*$ defined in (B.4),*

$$\begin{aligned}\left\|\Sigma_Z^{1/2}\left(A^\top\widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2 \lesssim_\theta &\left(1 + \frac{\delta_W}{\widehat{\eta}}\right)\left(\frac{K\wedge\widehat{r} + \log n}{n}\sigma^2 + \beta^\top A^+\Sigma_W A^{+\top}\beta\right)\\ &+ \left[\left(1 + \frac{\delta_W}{\widehat{\eta}}\right)\widehat{\psi} + \delta_W\right]\beta^\top(A^\top A)^{-1}\beta.\end{aligned}$$

Proof. Use identity (B.7) and the inequality $(x+y)^2 \le 2x^2 + 2y^2$ to find

$$\begin{aligned}&\left\|\Sigma_Z^{1/2}\left(A^\top\widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2\\ &\le 2\left\|\Sigma_Z^{1/2}[A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2 + 2\left\|\Sigma_Z^{1/2}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2.\end{aligned} \tag{B.13}$$

For the first term, since $\boldsymbol{Z}\in\mathbb{R}^{n\times K}$ has $\mathrm{rank}(\boldsymbol{Z}) = K$ on the event $\mathcal{E}^*$, we have

$$\begin{aligned}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+ - \boldsymbol{Z}^+ &= \boldsymbol{Z}^+\boldsymbol{Z}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+ - \boldsymbol{Z}^+ \qquad (\text{by } \boldsymbol{Z}^+\boldsymbol{Z} = \boldsymbol{I}_K \text{ on } \mathcal{E}^*)\\ &= \boldsymbol{Z}^+(\boldsymbol{X} - \boldsymbol{W})\widehat{B}(\boldsymbol{X}\widehat{B})^+ - \boldsymbol{Z}^+\\ &= -\boldsymbol{Z}^+P^\perp_{\boldsymbol{X}\widehat{B}} - \boldsymbol{Z}^+\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+,\end{aligned} \tag{B.14}$$

which yields

$$\begin{aligned}&\left\|\Sigma_Z^{1/2}[A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2\\ &\le 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 + 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2\\ &\lesssim \frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 + \frac{1}{n}\left\|\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2\\ &\lesssim \frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 + \delta_W\cdot\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2.\end{aligned} \tag{B.15}$$

We used $\|\Sigma_Z^{1/2}\boldsymbol{Z}^+\|_{\mathrm{op}} = \sigma_K^{-1}(\boldsymbol{Z}\Omega^{-1/2}) \lesssim 1/\sqrt{n}$ on $\mathcal{E}^*$ in the third line. The event $\mathcal{E}_{Z\beta}$ and (B.12) conclude

$$\begin{aligned}\left\|\Sigma_Z^{1/2}[A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2 \\ \lesssim \left(1+\frac{\delta_W}{\widehat{\eta}}\right)\left(\beta^\top A^+\Sigma_W A^{+\top}\beta + \widehat{\psi}\beta^\top(A^\top A)^{-1}\beta\right) + \delta_W\beta^\top(A^\top A)^{-1}\beta.\end{aligned} \tag{B.16}$$

For the second term in (B.13), we use that on $\mathcal{E}^*$ (in particular, $\mathcal{E}_{M'} \subset \mathcal{E}^*$),

$$\left\|\Sigma_Z^{1/2}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 \le 2\gamma_\varepsilon^2\sigma^2\left[2\|M'\|_{\mathrm{op}}\log n + \mathrm{tr}(M')\right]$$

Since $\mathrm{rank}(\Sigma_Z) = K$ and $\mathrm{rank}(\boldsymbol{X}\widehat{P}_{\widehat{B}}) = \widehat{r}$, we have

$$\mathrm{tr}(M') \le (K \wedge \widehat{r})\|M'\|_{\mathrm{op}}.$$

Moreover,

$$\begin{aligned}\|M'\|_{\mathrm{op}} = \left\|\Sigma_Z^{1/2}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2 &\le 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+P_{\boldsymbol{X}\widehat{B}}\right\|_{\mathrm{op}}^2 + 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2 \\ &\lesssim \frac{1}{n} + \delta_W\cdot\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\end{aligned}$$

by using (B.14) in the first line and $\mathcal{E}^*$ in the second line. Invoking (B.8) concludes that, on $\mathcal{E}^*$,

$$\left\|\Sigma_Z^{1/2}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 \lesssim \frac{(K\wedge\widehat{r}+\log n)\sigma^2}{n}\left(1+\frac{\delta_W}{\widehat{\eta}}\right). \tag{B.17}$$

Plugging (B.16) and (B.17) into (B.13) completes the proof. ■

Lemma 44. *Under conditions of Theorem 17, on the event $\mathcal{E}'_{\boldsymbol{W}}$ from (B.4),*

$$\frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 \le 8\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta + 2\widehat{\psi}\beta^\top(A^\top A)^{-1}\beta. \tag{B.18}$$

Proof. By $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$, one has

$$\begin{aligned}P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta &= P^\perp_{\boldsymbol{X}\widehat{B}}(\boldsymbol{X}A^{+\top} - \boldsymbol{W}A^{+\top})\beta \\ &= -P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta + P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}A^{+\top}\beta \\ &= -P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta + P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}\left(A^{+\top} - \widehat{B}G\right)\beta\end{aligned}$$

for any matrix $G \in \mathbb{R}^{q\times K}$. Choose

$$G = \widehat{B}^{+}A^{+\top} = \min_{G'} \left\| A^{+\top} - \widehat{B}G' \right\|_F$$

to obtain

$$P^{\perp}_{X\widehat{B}}\mathbf{Z}\beta = P^{\perp}_{X\widehat{B}}\boldsymbol{W}A^{+\top}\beta + P^{\perp}_{X\widehat{B}}\boldsymbol{X}P^{\perp}_{\widehat{B}}A^{+\top}\beta.$$

Then by the basic inequality $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned} \left\| P^{\perp}_{X\widehat{B}}\mathbf{Z}\beta \right\|^2 &\le 2\left\| P^{\perp}_{X\widehat{B}}\boldsymbol{W}A^{+\top}\beta \right\|^2 + 2\left\| P^{\perp}_{X\widehat{B}}\boldsymbol{X}P^{\perp}_{\widehat{B}}A^{+\top}\beta \right\|^2 \\ &\le 2\left\| P^{\perp}_{X\widehat{B}} \right\|^2_{\mathrm{op}} \left\| \boldsymbol{W}A^{+\top}\beta \right\|^2 + 2\left\| \boldsymbol{X}P^{\perp}_{\widehat{B}} \right\|^2_{\mathrm{op}} \left\| A^{+\top}\beta \right\|^2 \\ &\le 2\left\| \boldsymbol{W}A^{+\top}\beta \right\|^2 + 2n\widehat{\psi}\beta^{\top}(A^{\top}A)^{-1}\beta \end{aligned} \tag{B.19}$$

where we invoked the definition of $\widehat{\psi}$ in the last line. Invoke $\mathcal{E}'_{\boldsymbol{W}}$ from (B.4) to finish the proof. ■

B.2.2 Proofs for Section 2.3

Proof of Corollary 18

The corollary is an application of Theorem 17 with $\widehat{B} = \mathbf{U}_k$. Given any realization of $(\boldsymbol{X}, \boldsymbol{Y})$ and (possibly random) $k \in \{0, 1, \ldots, \mathrm{rank}(\boldsymbol{X})\}$, we may write the SVD of $\boldsymbol{X}$ as

$$\begin{aligned} \boldsymbol{X} = V\Delta\mathbf{U}^{\top} &= \sum_{1\le j\le k} \Delta_{jj}\boldsymbol{V}_{\bullet j}\mathbf{U}^{\top}_{\bullet j} + \sum_{j>k} \Delta_{jj}\boldsymbol{V}_{\bullet j}\mathbf{U}^{\top}_{\bullet j} \\ &:= \boldsymbol{V}_k\Delta_k\mathbf{U}^{\top}_k + \boldsymbol{V}_{(-k)}\Delta_{(-k)}\mathbf{U}^{\top}_{(-k)}. \end{aligned}$$

The diagonal matrix Δ contains the non-increasing singular values and $\mathbf{U}_k$ contains the corresponding k right-singular vectors. Consequently,

$$\begin{aligned}
&\operatorname{rank}(\boldsymbol{X}\mathbf{U}_k) = \operatorname{rank}(\boldsymbol{V}_k\Delta_k) = k, \\
&\sigma_1^2\left(\boldsymbol{X}P_{\mathbf{U}_k}^{\perp}\right) = \left\|\boldsymbol{X}\mathbf{U}_{(-k)}\mathbf{U}_{(-k)}^{\top}\right\|_{\mathrm{op}}^2 = \left\|\boldsymbol{V}_{(-k)}\Delta_{(-k)}\mathbf{U}_{(-k)}^{\top}\right\|_{\mathrm{op}}^2 = \sigma_{k+1}^2(\boldsymbol{X}) = n\widehat{\lambda}_{k+1}, \\
&\sigma_1^2(\boldsymbol{X}P_{\mathbf{U}_k}) = \sigma_1^2(\boldsymbol{V}_k\Delta_k\mathbf{U}_k^{\top}) = \sigma_k^2(\boldsymbol{X}) = n\widehat{\lambda}_k.
\end{aligned}$$

Invoke Theorem 17 with $\widehat{B} = \mathbf{U}_k$, $\widehat{r} = k$, $\widehat{\psi} = \widehat{\lambda}_{k+1}$ and $\widehat{\eta} = \widehat{\lambda}_k$ to conclude the proof.■

Proof of Corollary 19 & Remark 6

We first prove Corollary 19. From Corollary 18, it suffices to show $\mathbb{P}_\theta\{\widehat{s} \le K\} \ge 1 - c/n$, which is guaranteed by proving

$$\mathbb{P}_\theta\left\{\frac{1}{n}\sigma_{K+1}^2(\boldsymbol{X}) < C_0\delta_W\right\} \ge 1 - c/n.$$

By Weyl's inequality,

$$\sigma_{K+1}(\boldsymbol{X}) \le \sigma_{K+1}(\boldsymbol{Z}A^{\top}) + \sigma_1(\boldsymbol{W}) = \sigma_1(\boldsymbol{W}).$$

The result then follows by (2.11) and $C_0 > 1$. ■

To prove Remark 6, we will show

$$\mathbb{P}\left\{\widehat{\lambda}_K \gtrsim \lambda_k(A\Sigma_Z A^{\top}) - \delta_W\right\} \ge 1 - n^{-c}.$$

Note that Weyl's inequality yields

$$\sigma_k(\boldsymbol{X}) \ge \sigma_k(\boldsymbol{Z}A^{\top}) - \sigma_1(\boldsymbol{W}) \ge \sigma_K(\boldsymbol{Z}\Sigma_Z^{-1/2})\sigma_k(\Sigma_Z^{1/2}A^{\top}) - \sigma_1(\boldsymbol{W}).$$

We obtain the desired result by invoking $\mathcal{E}_Z$ from Lemma 41 and (2.11). ■

Proof of Proposition 20

We work on the event

$$\mathcal{E}''_{\boldsymbol{W}} := \left\{\sigma_1^2(\boldsymbol{W}) \le n\delta_W\right\} \cap \left\{c_1 \operatorname{tr}(\Sigma_W) \le \frac{1}{n}\|\boldsymbol{W}\|_F^2 \le C_1 \operatorname{tr}(\Sigma_W)\right\}$$

with δ_W defined in (2.12) and some constants $C_1 \ge c_1 > 0$, depending on γ_w. We have on the event $\mathcal{E}''_{\boldsymbol{W}}$,

$$\begin{aligned}
2\sigma_1^2(\boldsymbol{W})\frac{np}{\|\boldsymbol{W}\|_F^2} &\le 2n\delta_W \frac{np}{\|\boldsymbol{W}\|_F^2} \\
&\le \frac{2\delta_W}{c_1}\frac{np}{\operatorname{tr}(\Sigma_W)} && \text{by } \mathcal{E}''_{\boldsymbol{W}} \\
&= \frac{2c}{c_1}\left(\frac{np}{r_e(\Sigma_W)} + p\right) && \text{by (2.12)} \\
&\le \frac{2c}{c_1}\left(\frac{n \vee p}{c'} + p\right) && \text{by } r_e(\Sigma_W) \ge c'(n \wedge p) \\
&\le c_0(n+p) = \mu_n
\end{aligned}$$

by choosing any $c_0 \ge 2c(1 + 1/c')/c_1$. From Theorem 6 and Proposition 7 of [27] with $P = \boldsymbol{I}_n$, $E = \boldsymbol{W}$ and $m = p$, we deduce

$$\widetilde{s} \le K$$

on the event $\mathcal{E}''_{\boldsymbol{W}}$.

To prove the lower bound $\sigma_{\widetilde{s}}^2(\boldsymbol{X}) \gtrsim n\delta_W$, we notice that, on the event $\mathcal{E}''_{\boldsymbol{W}}$,

$$\sigma_{\widetilde{s}}^2(\boldsymbol{X}) \ge \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(\widetilde{s})}\|_F^2}{np - \mu_n\widetilde{s}} \ge \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np}. \tag{B.20}$$

The first inequality uses (2.7) in [27], while the second inequality uses $K \le \bar{K}$. Further invoking (3.8) in Proposition 7 of [27] yields

$$\frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np - \mu_n K} \ge \frac{\|\boldsymbol{W}\|_F^2}{np}.$$

Next, on the event $\mathcal{E}''_{\boldsymbol{W}}$, choosing $c_0 \geq 2c(1 + 1/c')/c_1$ in $\mu_n = c_0(n + p)$, we find

$$\begin{aligned}
\mu_n \frac{\|\boldsymbol{W}\|_F^2}{np} &\geq \mu_n c_1 \frac{\mathrm{tr}(\Sigma_W)}{p} \\
&\geq 2c\left(1 + \frac{1}{c'}\right)\frac{n+p}{p}\mathrm{tr}(\Sigma_W) \\
&\geq 2c\left(\mathrm{tr}(\Sigma_W) + \frac{1}{c'}\frac{n+p}{p} r_e(\Sigma_W)\|\Sigma_W\|_{\mathrm{op}}\right) \\
&\geq 2c\left(\mathrm{tr}(\Sigma_W) + (n \wedge p)\frac{n+p}{p}\|\Sigma_W\|_{\mathrm{op}}\right) \qquad \text{by } r_e(\Sigma_W) \geq c'(n \wedge p) \\
&\geq 2c\left(\mathrm{tr}(\Sigma_W) + n\|\Sigma_W\|_{\mathrm{op}}\right) \\
&= 2n\delta_W.
\end{aligned}$$

Hence, combining all three previous displays, we derive

$$\begin{aligned}
\sigma^2_{\widetilde{s}}(\boldsymbol{X}) &\geq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np} \\
&\geq \mu_n \frac{\|\boldsymbol{W}\|_F^2}{np}\frac{np - \mu_n K}{np} \\
&\geq n\delta_W \frac{np - \mu_n K}{np} \\
&\geq \frac{1}{1+\kappa} n\delta_W \qquad \text{by } K \leq \bar{K} \text{ and (2.21)}.
\end{aligned}$$

Next, we prove $\sigma^2_{\widetilde{s}+1}(\boldsymbol{X}) \lesssim \delta_W$. By (2.7) in [27] once again, we have

$$\sigma^2_{\widetilde{s}+1}(\boldsymbol{X}) \leq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(\widetilde{s}+1)}\|_F^2}{np - \mu_n(\widetilde{s}+1)}.$$

From (2.3) in Proposition 1 of [27], this inequality is equivalent to

$$\sigma^2_{\widetilde{s}+1}(\boldsymbol{X}) \leq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(\widetilde{s})}\|_F^2}{np - \mu_n \widetilde{s}}.$$

Since $\widetilde{s} \le K$ on $\mathcal{E}''_{\boldsymbol{W}}$, we have

$$\begin{aligned}
\sigma^2_{\widetilde{s}+1}(\boldsymbol{X}) &\le \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np - \mu_n K} \\
&\le \mu_n \frac{np}{np - \mu_n K} \frac{\|\boldsymbol{W}\|_F^2}{np} && \text{by (3.8) of Proposition 7 in [27]} \\
&\le (1+\kappa)\mu_n \frac{\|\boldsymbol{W}\|_F^2}{np} && \text{by (2.21)} \\
&\le (1+\kappa)c_0 C_1 (n+p)\frac{\mathrm{tr}(\Sigma_W)}{p} && \text{by } \mathcal{E}''_{\boldsymbol{W}} \text{ and } \mu_n = c_0(n+p) \\
&\le \frac{(1+\kappa)c_0 C_1}{c} n\delta_W && \text{by } \mathrm{tr}(\Sigma_W) \le p\|\Sigma_W\|_{\mathrm{op}}.
\end{aligned}$$

It remains to prove $1 - \mathbb{P}(\mathcal{E}''_{\boldsymbol{W}}) \lesssim 1/n$. First note that

$$\frac{1}{n}\|\boldsymbol{W}\|_F^2 = \sum_{j=1}^{p} \frac{1}{n} \boldsymbol{W}_{\cdot j}^\top \boldsymbol{W}_{\cdot j}.$$

By invoking Lemma 49 for fixed $j \in [p]$ and some absolute constant c, the inequality

$$\left| \frac{1}{n} \boldsymbol{W}_{\cdot j}^\top \boldsymbol{W}_{\cdot j} - [\Sigma_W]_{jj} \right| \le c\gamma_w^2 [\Sigma_W]_{jj} \sqrt{\frac{\log p}{n}}$$

holds with probability at least $1 - 2(p \vee n)^{-2}$. Apply the union bound over $1 \le j \le p$, invoke $\log p \le Cn$ for sufficiently large C, and conclude

$$\mathbb{P}\left\{ c(\gamma_w)\,\mathrm{tr}(\Sigma_W) \le \frac{1}{n}\|\boldsymbol{W}\|_F^2 \le C(\gamma_w)\,\mathrm{tr}(\Sigma_W) \right\} \ge 1 - 2(p \vee n)^{-1}.$$

Finally, Lemma 47 shows that $\mathbb{P}\{\sigma_1^2(\boldsymbol{W}) \le n\delta_W\} \ge 1 - e^{-n}$, taking c in δ_W large enough. ■

B.2.3 Proofs for Section 2.4

Proof of Corollary 22

By Theorem 5.39 of [90], $\sigma_p^2(\boldsymbol{X}\Sigma_X^{-1/2}) \gtrsim n$ with probability at least $1 - cn^{-1}$, where we use that $\boldsymbol{X}\Sigma_X^{-1/2}$ has independent sub-Gaussian rows with sub-Gaussian constant

bounded by an absolute constant, which is implied by the sub-Gaussianity of Z and W, and that $p \log n \lesssim n$. Thus, with the same probability,

$$\sigma_p^2(\boldsymbol{X}) \geq \lambda_p(\Sigma_X)\sigma_p^2(\boldsymbol{X}\Sigma_X^{-1/2}) \geq \lambda_p(\Sigma_W)\sigma_p^2(\boldsymbol{X}\Sigma_X^{-1/2}) \gtrsim \lambda_p(\Sigma_W)n.$$

Corollary 22 then follows from Theorem 17 with $\widehat{\psi} = 0$, $\widehat{\eta} \gtrsim \lambda_p(\Sigma_W)$, and $\widehat{r} \leq p$. ■

Proof of Corollary 23

Under conditions of Corollary 23, [33] proves that

$$\mathbb{P}\left\{\sigma_n^2(\boldsymbol{X}) \gtrsim \operatorname{tr}(\Sigma_W)\right\} \geq 1 - cn^{-1}.$$

We thus have $r = n$, $\widehat{\psi} = 0$, and $\widehat{\eta} \gtrsim \operatorname{tr}(\Sigma_W)/n$. Further noting that

$$\delta_W = \|\Sigma_W\|_{\mathrm{op}}\left(1 + \frac{r_e(\Sigma_W)}{n}\right) \asymp \frac{\operatorname{tr}(\Sigma_W)}{n},$$

such that $\delta_W/\widehat{\eta} \asymp 1$, we conclude

$$\begin{aligned}\mathbb{R}^*(\boldsymbol{I}_p) - \sigma^2 &\lesssim \frac{K + \log n}{n}\sigma^2 + \frac{n}{r_e(\Sigma_W)}\sigma^2 + \frac{\operatorname{tr}(\Sigma_W)}{n}\beta^\top(A^\top A)^{-1}\beta \\ &\lesssim \frac{K + \log n}{n}\sigma^2 + \frac{n}{r_e(\Sigma_W)}\sigma^2 + \frac{r_e(\Sigma_W)}{n}\|\Sigma_W\|_{\mathrm{op}}\,\beta^\top(A^\top A)^{-1}\beta.\end{aligned}$$

■

Proof of Theorem 24

Instead of directly applying Theorem 17, we slightly modify the proofs of Theorem 17 to obtain a sharp result for $\mathbb{R}(\widehat{A})$.

From the proof of Theorem 17, display (B.2) gives

$$\mathbb{R}(\widehat{A}) - \sigma^2 \leq \left\|\Sigma_Z^{1/2}\left(A^\top\widehat{\alpha}_{\widehat{A}} - \beta\right)\right\|^2 + \|\Sigma_W\|_{\mathrm{op}}\left\|\widehat{\alpha}_{\widehat{A}}\right\|^2.$$

We then point out the modifications of the proof of Lemmas 42 and 43. Recall $\widehat{A} \in \mathbb{R}^{p\times\widehat{K}}$. We work on the event $\mathcal{E}^*$ defined in the proof of Theorem 17 intersected with the event that $\widehat{K} = K$ and

$$\|\widehat{A} - A\|_{\mathrm{op}}^2 \le \|\widehat{A} - A\|_F^2 \lesssim \|A_J\|_0 \frac{\log(p \vee n)}{n}.$$

The last two events holds with probability at least $1 - c(p \vee n)^{-1}$ for some constant $c > 0$ [26]. In display (B.11) of Lemma 42 for bounding $\|\widehat{\alpha}_{\widehat{A}}\|^2$, we use

$$\begin{aligned}
\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2 &\le 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+A^{+\top}\beta\right\|^2 + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}^{\perp}A^{+\top}\beta\right\|^2 \\
&\quad + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{W}A^{+\top}\beta\right\|^2 \\
&\le 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+\right\|_{\mathrm{op}}^2\left\|A^{+\top}\beta\right\|^2 + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\left\|\boldsymbol{X}P_{\widehat{B}}^{\perp}A^{+\top}\beta\right\|^2 \\
&\quad + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2.
\end{aligned}$$

We change the way to bound the second term on the right hand side. Specifically, set $\widehat{B} = \widehat{A}$ and use $(a+b)^2 \le 2a^2 + 2b^2$ twice to obtain

$$\begin{aligned}
\left\|\boldsymbol{X}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 &\le 2\left\|\boldsymbol{Z}AP_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 + 2\left\|\boldsymbol{W}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 \\
&\le 2\left\|\boldsymbol{Z}\Omega^{1/2}\right\|_{\mathrm{op}}^2\left\|\Sigma_Z^{1/2}(A-\widehat{A})^{\top}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 \qquad (\text{by } \widehat{A}^{\top}P_{\widehat{A}}^{\perp} = 0) \\
&\quad + 4\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 + 4\left\|\boldsymbol{W}P_{\widehat{A}}A^{+\top}\beta\right\|^2 \qquad (\text{by } P_{\widehat{A}}^{\perp} = \boldsymbol{I}_p - P_{\widehat{A}}).
\end{aligned}$$

By $\mathcal{E}_Z$, $\mathcal{E}'_W$ and Lemma 45, after a bit algebra, we conclude

$$\begin{aligned}
\frac{1}{n}\left\|\boldsymbol{X}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 &\lesssim \left(\|A_J\|_0\frac{\log(p\vee n)}{n} + \delta_{W,J}\right)\beta^T(A^{\top}A)^{-1}\beta + \beta^{\top}A^{+}\Sigma_W A^{+\top}\beta \\
&\lesssim \left(\|A_J\|_0\frac{\log(p\vee n)}{n} + \|\Sigma_W\|_{\mathrm{op}}\right)\beta^T(A^{\top}A)^{-1}\beta + \beta^{\top}A^{+}\Sigma_W A^{+\top}\beta. \qquad \text{(B.21)}
\end{aligned}$$

with probability at least $1 - cn^{-1}$. In the last step, we used the fact that $\|\Sigma_W\|_{\mathrm{op}}$ is bounded and $\|A_J\|_{\ell_0/\ell_2} \le \|A_J\|_0$. Together with the proofs of Lemma 42, one can deduce that

$$\|\widehat{\alpha}_{\widehat{A}}\|^2 \lesssim \frac{(K+\log n)\sigma^2}{n\widehat{\eta}} + \beta^{\top}(A^{\top}A)^{-1}\beta + \widehat{\eta}^{-1}\left(\widehat{\psi}\beta^{\top}(A^{\top}A)^{-1}\beta + \beta^{\top}A^{+}\Sigma_W A^{+\top}\beta\right).$$

where

$$\widehat{\psi} \lesssim \|\Sigma_W\|_{\mathrm{op}} + \|A_J\|_0 \frac{\log(p \vee n)}{n}.$$

To bound $\|\Sigma_Z^{1/2}(A^\top \widehat{\alpha}_{\widehat{A}} - \beta)\|^2$, we modify two places in the proof of Lemma 43. Display (B.15) is bounded by

$$\begin{aligned}\left\|\Sigma_Z^{1/2}[A^\top \widehat{A}(\boldsymbol{X}\widehat{A})^+ \boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2 &\lesssim \frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{A}}\boldsymbol{Z}\beta\right\|^2 + \frac{1}{n}\left\|\boldsymbol{W}\widehat{A}(\boldsymbol{X}\widehat{A})^+\boldsymbol{Z}\beta\right\|^2 \\ &\lesssim \frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{A}}\boldsymbol{Z}\beta\right\|^2 + \frac{1}{n}\left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\mathrm{op}}^2\left\|\widehat{A}(\boldsymbol{X}\widehat{A})^+\boldsymbol{Z}\beta\right\|^2\end{aligned}$$

where we will invoke Lemma 45. For the first term of the right hand side, by (B.19), we have

$$\begin{aligned}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 &\le 2\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta\right\|^2 + 2\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}P^\perp_{\widehat{B}}A^{+\top}\beta\right\|^2 \\ &\le 2\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 + 2\left\|\boldsymbol{X}P^\perp_{\widehat{B}}A^{+\top}\beta\right\|^2\end{aligned}$$

which can be further bounded by using (B.21) and invoking the event $\mathcal{E}'_{\boldsymbol{W}}$. Collecting all these ingredients, we conclude

$$\begin{aligned}\left\|\Sigma_Z^{1/2}\left(A^\top\widehat{\alpha}_{\widehat{A}} - \beta\right)\right\|^2 \lesssim \left(1 + \frac{\delta_{W,J}}{\widehat{\eta}}\right)\left(\frac{K + \log n}{n}\sigma^2 + \beta^\top A^+ \Sigma_W A^{+\top}\beta\right) \\ + \left[\left(1 + \frac{\delta_{W,J}}{\widehat{\eta}}\right)\widehat{\psi} + \delta_{W,J}\right]\beta^\top(A^\top A)^{-1}\beta.\end{aligned}$$

It then remains to lower bound $\widehat{\eta}$ by bounding $\sigma_K(\boldsymbol{X}P_{\widehat{A}})$ from below. By Weyl's inequality, $\mathrm{rank}(\widehat{A}) = K$, we have

$$\begin{aligned}\sigma_K\left(\boldsymbol{X}P_{\widehat{A}}A(A^\top A)^{-1/2}\right) &\ge \sigma_K\left(\boldsymbol{X}A(A^\top A)^{-1/2}\right) - \left\|\boldsymbol{X}P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} \\ &\ge \sigma_K\left(\boldsymbol{X}AN^{-1/2}N^{1/2}(A^\top A)^{-1/2}\right) - \left\|\boldsymbol{X}P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} \\ &\ge \sigma_K\left(\boldsymbol{X}AN^{-1/2}\right)\sigma_K\left(N^{1/2}(A^\top A)^{-1/2}\right) - \left\|\boldsymbol{X}P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}.\end{aligned}$$

by writing $N = A^\top \Sigma A$. To lower bound $\sigma_K\left(\boldsymbol{X}AN^{-1/2}\right)$, using Weyl's inequality

again and invoking Lemma 48 yield

$$\begin{aligned}
&\lambda_K\left(N^{-1/2}A^\top\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}AN^{-1/2}\right)\\
&\gtrsim \lambda_K\left(N^{-1/2}A^\top\Sigma AN^{-1/2}\right)-\left\|N^{-1/2}A^\top\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}-\Sigma\right)AN^{-1/2}\right\|_{\mathrm{op}}\\
&\gtrsim 1-\sqrt{\frac{K\log n}{n}}-\frac{K\log n}{n}\gtrsim 1
\end{aligned}$$

with probability at least $1-cn^{-C}$. On the other hand, by $\boldsymbol{X}=\boldsymbol{Z}A^\top+\boldsymbol{W}$,

$$\begin{aligned}
\left\|\boldsymbol{X}P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} &\le \left\|\boldsymbol{Z}A^\top P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}+\left\|\boldsymbol{W}P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}\\
&\le \left\|\boldsymbol{Z}(A-\widehat{A})^\top\right\|_{\mathrm{op}}+\left\|\boldsymbol{W}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}+\left\|\boldsymbol{W}P_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}\\
&\le \left\|\boldsymbol{Z}\Omega^{1/2}\right\|_{\mathrm{op}}\sigma_1(\Sigma_Z)\left\|(A-\widehat{A})^\top\right\|_{\mathrm{op}}+\left\|\boldsymbol{W}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}+\left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\mathrm{op}}.
\end{aligned}$$

By $\mathcal{E}_{\boldsymbol{Z}}$ and Lemmas 45 and 47, we have

$$\frac{1}{n}\left\|\boldsymbol{X}P^\perp_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}}\lesssim \delta_{W,J}+\frac{\|A_J\|_0\log(p\vee n)}{n}\lesssim \|\Sigma_W\|_{\mathrm{op}}+\frac{\|A_J\|_0\log(p\vee n)}{n}$$

with probability at least $1-cn^{-1}$. Provided that

$$\lambda_K(A\Sigma_ZA^\top)\ge C\left(\|\Sigma_W\|_{\mathrm{op}}+\frac{\|A_J\|_0\log(p\vee n)}{n}\right)$$

for sufficiently small constant $C>0$, we then conclude that

$$\sigma_K^2\left(\boldsymbol{X}P_{\widehat{A}}A(A^\top A)^{-1/2}\right)\gtrsim n\lambda_K(A\Sigma_ZA^\top)$$

from noting $\sigma_K^2\left(N^{1/2}(A^\top A)^{-1/2}\right)=\lambda_K(A\Sigma_ZA^\top)$. This concludes $\widehat{\eta}\gtrsim\lambda_K(A\Sigma_ZA^\top)$. The result then follows by collecting terms. ■

The following lemma provides upper bounds for the operator norm of $\boldsymbol{W}P_{\widehat{A}}$. Recall that $\|A_J\|_{\ell_0/\ell_2}=\sum_{j\in J}1_{\{\|A_{j\bullet}\|_2\neq 0\}}$.

Lemma 45. *Under conditions of Theorem 24, with probability at least* $1-c(p\vee n)^{-1}$, *one has*

$$\frac{1}{n}\left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\mathrm{op}}^2\lesssim\|\Sigma_W\|_{op}\left(1+\frac{\|A_J\|_{\ell_0/\ell_2}}{n}\right):=\delta_{W,J}.$$

Proof. We work on the event $\widehat{K} = K$ and $\widehat{A}_I = A_I$ which holds with probability at least $1 - c(p \vee n)^{-c'}$ [26]. Then

$$\left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\text{op}} = \left\|\boldsymbol{W}\widehat{A}\widehat{A}^{+}\right\|_{\text{op}} \leq \left\|\boldsymbol{W}_{\cdot I}A_I\widehat{A}^{+}\right\|_{\text{op}} + \left\|\boldsymbol{W}_{\cdot J}\widehat{A}_J\widehat{A}^{+}\right\|_{\text{op}}$$
$$\leq \left\|\boldsymbol{W}_{\cdot I}A_I(A_I^{\top}A_I)^{-1/2}\right\|_{\text{op}}\left\|(A_I^{\top}A_I)^{1/2}\widehat{A}^{+}\right\|_{\text{op}} + \|\boldsymbol{W}_{\cdot J}\|_{\text{op}}\left\|\widehat{A}_J\widehat{A}^{+}\right\|_{\text{op}}.$$

Since

$$\left\|(A_I^{\top}A_I)^{1/2}\widehat{A}^{+}\right\|_{\text{op}}^2 = \left\|(A_I^{\top}A_I)^{1/2}(\widehat{A}^{\top}\widehat{A})^{-1}(A_I^{\top}A_I)^{1/2}\right\|_{\text{op}} \leq 1$$

by noting $\widehat{A}^{\top}\widehat{A} = A_I^{\top}A_I + \widehat{A}_J^{\top}\widehat{A}_J$, and similar arguments yield

$$\left\|\widehat{A}_J\widehat{A}^{+}\right\|_{\text{op}}^2 = \left\|\widehat{A}_J(\widehat{A}^{\top}\widehat{A})^{-1}\widehat{A}_J^{\top}\right\|_{\text{op}} = \left\|(\widehat{A}^{\top}\widehat{A})^{-1/2}\widehat{A}_J^{\top}\widehat{A}_J(\widehat{A}^{\top}\widehat{A})^{-1/2}\right\|_{\text{op}} \leq 1,$$

invoking Lemma 47 to bound $\|\boldsymbol{W}_{\cdot I}A_I(A_I^{\top}A_I)^{-1/2}\|_{\text{op}}$ and $\|W_{\cdot J}\|_{\text{op}}$ gives

$$\frac{1}{n}\left\|\boldsymbol{W}_{\cdot I}A_I(A_I^{\top}A_I)^{-1/2}\right\|_{\text{op}}^2 \lesssim \|\Psi_{II}\|_{\text{op}} + \frac{\text{tr}(\Psi_{II})}{n},$$
$$\frac{1}{n}\|\boldsymbol{W}_{\cdot J}\|_{\text{op}}^2 \lesssim \|[\Sigma_W]_{JJ}\|_{\text{op}} + \frac{\text{tr}([\Sigma_W]_{JJ})}{n} \leq \delta_{W,J},$$

with probability at least $1 - 2e^{-n}$, where

$$\Psi_{II} = (A_I^{\top}A_I)^{-1/2}A_I^{\top}[\Sigma_W]_{II}A_I(A_I^{\top}A_I)^{-1/2}.$$

The result then follows by using $\|\Psi_{II}\|_{\text{op}} \leq \|[\Sigma_W]_{II}\|_{\text{op}}$, $\text{tr}(\Psi_{II}) \leq K\|\Psi_{II}\|_{\text{op}} \leq K\|[\Sigma_W]_{II}\|_{\text{op}}$ and $K\log n \lesssim n$. ■

B.2.4 Proof of Theorem 25 in Section 2.5

For any $\alpha \in \mathbb{R}^p$, let

$$\widehat{\mathbb{R}}(\alpha) = \frac{2}{n}\sum_{i\in D_1}[Y_i - X_i^{\top}\alpha]^2$$

so that for all $m \in [M]$, by the definition of $\widehat{m}$, $\widehat{S}(\widehat{\alpha}) \le \widehat{S}(\widehat{\alpha}_m)$. Also let

$$\widehat{S}(\alpha) = \frac{2}{n}\sum_{i\in D_1}[Z_i^\top\beta - X_i^\top\alpha]^2.$$

Finally, for any fixed or random α define

$$S(\alpha) = \mathbb{E}_{(Z_*,X_*)}(Z_*^\top\beta - X_*^\top\alpha)^2, \qquad \mathbb{R}(\alpha) = S(\alpha) + \sigma^2,$$

where the expectation is over (Z_*, X_*) that are independent of α.

We have

$$\begin{aligned}
S(\widehat{\alpha}) &= \mathbb{R}(\widehat{\alpha}) - \sigma^2 \\
&= (1+a)[\widehat{\mathbb{R}}(\widehat{\alpha}) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] + [\mathbb{R}(\widehat{\alpha}) - (1+a)\widehat{\mathbb{R}}(\widehat{\alpha}) - (\sigma^2 - (1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i^2)].
\end{aligned}$$

Using $\widehat{\mathbb{R}}(\widehat{\alpha}) \le \widehat{\mathbb{R}}(\widehat{\alpha}_m)$ in the first term of the above, we have for any $m \in [M]$,

$$\begin{aligned}
S(\widehat{\alpha}) &\le (1+a)[\widehat{\mathbb{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] \\
&\quad + \max_m[\mathbb{R}(\widehat{\alpha}_m) - (1+a)\widehat{\mathbb{R}}(\widehat{\alpha}_m) - (\sigma^2 - (1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i^2)] \\
&= (1+a)[\widehat{\mathbb{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] \\
&\quad + \max_m[S(\widehat{\alpha}_m) - (1+a)\widehat{S}(\widehat{\alpha}_m) + 2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta)] \\
&\le (1+a)[\widehat{\mathbb{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] + \max_m[S(\widehat{\alpha}_m) - (1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m)] \\
&\quad + \max_m[2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)]. \qquad \text{(B.22)}
\end{aligned}$$

The first term in the above can be further re-written as

$$\begin{aligned}\widehat{\mathbb{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2 &= (1+a)S(\widehat{\alpha}_m) + [\widehat{\mathbb{R}}(\alpha_m) - (1+a)S(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}\varepsilon_i^2] \\ &= (1+a)S(\widehat{\alpha}_m) + [\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m) + \frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m)] \\ &\le (1+a)S(\widehat{\alpha}_m) + \max_m[(1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m)] \\ &\quad + \max_m[\frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)].\end{aligned}$$

Using this result in (B.22), we find that for any $m \in [M]$,

$$\begin{aligned}S(\widehat{\alpha}) &\le (1+a)^2 S(\widehat{\alpha}_m) \\ &\quad + (1+a)\max_m[(1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m)] \\ &\quad + (1+a)\max_m[\frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)] \\ &\quad + \max_m[S(\widehat{\alpha}_m) - (1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m)] \\ &\quad + \max_m[2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)] \\ &=: (1+a)^2 S(\widehat{\alpha}_m) + (1+a)T_1 + (1+a)T_2 + T_3 + T_4. \end{aligned} \tag{B.23}$$

Below we prove that

$$\mathbb{P}_\theta\left((1+a)T_1 + T_3 \le c_1\frac{(2+a)^3}{a}\cdot\frac{\max_m S(\widehat{\alpha}_m)\log(nM)}{n}\right) \ge 1 - c_1' n^{-1}, \tag{B.24}$$

and

$$\mathbb{P}_\theta\left\{(1+a)T_2 + T_4 \le c_2\frac{(1+a)^3}{a}\sigma^2\frac{\log(nM)}{n}\right\} \ge 1 - c_2' n^{-1}, \tag{B.25}$$

where c_1 and c_2 depend only on $\gamma_z, \gamma_w, \gamma_\varepsilon$ from Definition 2.2.1, and $c_1, c_2 > 0$ are absolute constants. The final result follows from taking a minimum over m in (B.23) and combining (B.24) and (B.25) with a union bound.

Bounding T_1 and T_3: Since $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_2$ are independent of $\{X_i : i \in D_1\}$, we will prove (B.24) for the case when $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_2$ are non-random without loss of generality.

We first consider T_3. For all $t, b > 0$, the following holds:

$$S - \widehat{S} \leq \sqrt{t}\sqrt{S} \quad \Rightarrow \quad S \leq (1+b)\widehat{S} + t\frac{1+b}{b}, \tag{B.26}$$

where we write $S = S(\widehat{\alpha}_m)$ and $\widehat{S} = \widehat{S}(\widehat{\alpha}_m)$. To prove this, suppose the left hand side holds true and consider the cases $\sqrt{S} \leq \frac{1+b}{b}\sqrt{t}$, which implies $S \leq \widehat{S} + t\frac{1+b}{b}$, and $\sqrt{S} > \frac{1+b}{b}\sqrt{t}$, which implies $S \leq \widehat{S} + \frac{b}{1+b}S$ and thus $S \leq (1+b)\widehat{S}$. Thus,

$$\begin{aligned}
\mathbb{P}_\theta\left(T_3 > t\frac{1+a/2}{a/2}\right) &\leq M \max_m \mathbb{P}_\theta\left(S(\widehat{\alpha}_m) - (1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m) > t\frac{1+a/2}{a/2}\right) \\
&\leq M \max_m \mathbb{P}_\theta\left(\frac{S(\widehat{\alpha}_m) - \widehat{S}(\widehat{\alpha}_m)}{\sqrt{S(\widehat{\alpha}_m)}} > \sqrt{t}\right) \qquad \text{(by (B.26))} \\
&\leq M \max_m \mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_1}[\mathbb{E}[g_i(m)] - g_i(m)]\right| > \sqrt{t}\right), \qquad \text{(B.27)}
\end{aligned}$$

where we let $g_i(m) := (Z_i^\top\beta - X_i^\top\widehat{\alpha}_m)^2/\sqrt{S(\widehat{\alpha}_m)}$ in the last step. Recalling that for any random variable U, $\|U^2\|_{\psi_1} = \|U\|_{\psi_2}^2$, and using the assumption that $\widehat{\alpha}_m$ is a fixed vector, we find

$$\begin{aligned}
&\|(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m)^2\|_{\psi_1} \\
&= \|Z_i^\top\beta - X_i^\top\widehat{\alpha}_m\|_{\psi_2}^2 \\
&\leq \|Z_i^\top\beta - Z_i^\top A^\top\widehat{\alpha}_m\|_{\psi_2}^2 + \|W_i^\top\widehat{\alpha}_m\|_{\psi_2}^2 && \text{(since } X_i = AZ_i + W_i\text{)} \\
&= \|(\Sigma_Z^{-1/2}Z_i)^\top(\Sigma_Z^{1/2}[\beta - A^\top\widehat{\alpha}_m])\|_{\psi_2}^2 + \|(\Sigma_W^{-1/2}W)^\top(\Sigma_W^{1/2}\widehat{\alpha}_m)\|_{\psi_2}^2 \\
&= \|\Sigma_Z^{1/2}(\beta - A^\top\widehat{\alpha}_m)\|^2\|(\Sigma_Z^{-1/2}Z_i)^\top u)\|_{\psi_2}^2 && \text{(with } \|u\| = \|v\| = 1\text{)} \\
&\qquad + \|\Sigma_W^{1/2}\widehat{\alpha}_m\|^2\|(\Sigma_W^{-1/2}W)^\top v\|_{\psi_2}^2 \\
&\leq c_1\|\Sigma_Z^{1/2}(\beta - A^\top\widehat{\alpha}_m)\|^2 + c_1\|\Sigma_W^{1/2}\widehat{\alpha}_m\|^2 && \text{(by Definition (2.2.1))} \\
&= c_1 S(\widehat{\alpha}_m),
\end{aligned}$$

where $c_1 = c_1(\gamma_z, \gamma_w)$. Thus,

$$\|\mathbb{E}g_i(m) - g_i(m)\|_{\psi_1} \lesssim \|g_i(m)\|_{\psi_1} \leq c_1 \sqrt{S(\widehat{\alpha}_m)},$$

so by Bernstein's inequality [90],

$$\mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_1}[\mathbb{E}[g_i(m)] - g_i(m)]\right| > \sqrt{t}\right) \leq 2\exp\left(-n\left(\frac{t}{c_1 S(\widehat{\alpha}_m)} \wedge \sqrt{\frac{t}{c_1 S(\widehat{\alpha}_m)}}\right)\right). \quad \text{(B.28)}$$

Choosing $t = c_1 \max_m S(\widehat{\alpha}_m)\log(nM)/n$, and combining with (B.27), for $\log(M) < cn$,

$$\mathbb{P}_\theta\left(T_3 > \frac{1 + a/2}{a/2} \cdot c_1 \frac{\max_m S(\widehat{\alpha}_m)\log(nM)}{n}\right) \leq 2/n. \quad \text{(B.29)}$$

We next consider T_1. For $t, b > 0$, we have

$$\widehat{S} - S \leq \sqrt{t}\sqrt{S} \quad \Rightarrow \quad \widehat{S} \leq \left(1 + \frac{b}{1+b}\right)S + t\frac{1+b}{b}.$$

To prove this, suppose the left hand side holds and consider the cases $\sqrt{S} \leq \frac{1+b}{b}\sqrt{t}$, which implies $\widehat{S} \leq S + \frac{1+b}{b}t$, and $\sqrt{S} > \frac{1+b}{b}\sqrt{t}$, which implies $\widehat{S} \leq [1 + b/(1+b)]S$. Multiplying the right hand inequality by $(1+b)$, and choosing $b = a/2$, we find

$$\left(1 + \frac{a}{2}\right)\widehat{S} - (1+a)S > t\frac{(1+a/2)^2}{a/2} \quad \Rightarrow \quad \widehat{S} - S > \sqrt{t}\sqrt{S} \quad \text{(B.30)}$$

Recalling

$$T_1 = \max_m[(1 + \frac{a}{2})\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m)],$$

an application of (B.30) gives

$$\begin{aligned}\mathbb{P}_\theta\left(T_1 > t\frac{(1+a/2)^2}{a/2}\right) &\leq M\max_m \mathbb{P}_\theta(\widehat{S}(\widehat{\alpha}_m) - S(\widehat{\alpha}_m) > \sqrt{t}\sqrt{S}) \\ &\leq M\max_m \mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_1}[\mathbb{E}[g_i(m)] - g_i(m)]\right| > \sqrt{t}\right)\end{aligned}$$

Choosing $t = c_1 \max_m S(\widehat{\alpha}_m)\log(nM)/n$ and applying (B.28) with $\log(M) < cn$, we conclude

$$\mathbb{P}_\theta\left(T_1 > \frac{(1+a/2)^2}{a/2} \cdot c_1 \frac{\max_m S(\widehat{\alpha}_m)\log(nM)}{n}\right) \leq 2/n. \quad \text{(B.31)}$$

Combining (B.29) and (B.31) with a union bound and some algebra proves (B.24).

Bounding T_2 and T_4: For each $i \in D_1$, define $h_i(m) = (Z_i^\top \beta - X_i^\top \widehat{\alpha}_m)/[\widehat{S}(\widehat{\alpha}_m)]^{1/2}$. Using the inequality $2|xy| \le x^2/c + cy^2$ for $c > 0$, we have that

$$\begin{aligned}\frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m) &= 2[\widehat{S}(\widehat{\alpha}_m)]^{1/2}\frac{2}{n}\sum_{i\in D_1}\varepsilon_i h_i(m) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)\\ &\le 2[\widehat{S}(\widehat{\alpha}_m)]^{1/2}\left|\frac{2}{n}\sum_{i\in D_1}\varepsilon_i h_i(m)\right| - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)\\ &\le \frac{2}{a}\left|\frac{2}{n}\sum_{i\in D_1}\varepsilon_i h_i(m)\right|^2\end{aligned}$$

Similarly,

$$2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m) \le \frac{2(1+a)^2}{a}\left|\frac{2}{n}\sum_{i\in D_1}\varepsilon_i h_i(m)\right|^2.$$

Thus,

$$T_2 + T_4 \lesssim \max_m \frac{(1+a)^2}{a}\left|\frac{2}{n}\sum_{i\in D_1}\varepsilon_i h_i(m)\right|^2,$$

so

$$\mathbb{P}_\theta\left(T_2 + T_4 \ge t\frac{(1+a)^2}{a}\right) \le M\max_m \mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_2}\varepsilon_i h_i(m)\right| \ge \sqrt{t}\right)$$

Since $\{\varepsilon_i\}_{i\in D_1}$ is independent of $(Z_i, X_i)_{i\in D_2}$, $\mathbb{E}[\varepsilon_i h_i(m)] = 0$ for all $i \in D_2$. Furthermore, $\|\varepsilon_i\|_{\psi_2} \lesssim \sigma$ and $|h_i(m)|$ is bounded by 1, so $\|\varepsilon_i h_i(m)\|_{\psi_2} \le \sigma/c_2$, where $c_2 = c_2(\gamma_\varepsilon)$. Thus by Hoeffding's inequality [90],

$$\mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_2}\varepsilon_i h_i(m)\right| \ge \sqrt{t}\right) \le 2\exp(-c_2 tn/\sigma^2).$$

Choosing $t = \sigma^2\log(nM)/(c_2 n)$ completes the proof of (B.25). ■

B.3 Auxiliary Lemmas

The following lemma is used in our analysis. The tail inequality is for a quadratic form of sub-Gaussian random vectors. It is a slightly simplified version of Lemma 30 in [54].

Lemma 46. *Let $\xi \in \mathbb{R}^d$ be a γ_ξ sub-Gaussian random vector. For all symmetric positive semi-definite matrices H, and all $t \geq 0$,*

$$\mathbb{P}\left\{\xi^\top H\xi > \gamma_\xi^2\left(\sqrt{\mathrm{tr}(H)} + \sqrt{2\|H\|_{op}t}\right)^2\right\} \leq e^{-t}.$$

Proof. From Lemma 8 in [54], one has

$$\mathbb{P}\left\{\xi^\top H\xi > \gamma_\xi^2\left(\mathrm{tr}(H) + 2\sqrt{\mathrm{tr}(H^2)t} + 2\|H\|_{\mathrm{op}}t\right)\right\} \leq e^{-t},$$

for all $t \geq 0$. The result then follows from $\mathrm{tr}(H^2) \leq \|H\|_{\mathrm{op}}\mathrm{tr}(H)$. ■

The following lemma provides an upper bound on the operator norm of $\boldsymbol{G}H\boldsymbol{G}^\top$ where $\boldsymbol{G} \in \mathbb{R}^{n\times d}$ is a random matrix and its rows are independent sub-Gaussian random vectors. It differs from [33, Theorem 10] in the sense that independence across columns of $\boldsymbol{G}$ is not required.

Lemma 47. *Let $\boldsymbol{G}$ be n by d matrix whose rows are independent γ sub-Gaussian random vectors with identity covariance matrix. Then for all symmetric positive semi-definite matrices H,*

$$\mathbb{P}\left\{\frac{1}{n}\|\boldsymbol{G}H\boldsymbol{G}^\top\|_{\mathrm{op}} \leq \gamma^2\left(\sqrt{\frac{\mathrm{tr}(H)}{n}} + \sqrt{6\|H\|_{op}}\right)^2\right\} \geq 1 - e^{-n}$$

Proof. By definition and the property of the 1/2-net $\mathbb{N}$,

$$\|\boldsymbol{G}H\boldsymbol{G}^\top\|_{\mathrm{op}} = \sup_{u\in\mathcal{S}^{n-1}} u^\top\boldsymbol{G}H\boldsymbol{G}^\top u \leq 2\sup_{u\in\mathbb{N}} u^\top\boldsymbol{G}H\boldsymbol{G}^\top u.$$

For fixed $u \in \mathbb{N}$, since $\boldsymbol{G}^\top u$ is a γ sub-Gaussian random vector, an application of Lemma 46 with $\xi = \boldsymbol{G}^\top u$, $\gamma_\xi = \gamma$ and $H = H$ yields

$$\mathbb{P}\left\{u^\top \boldsymbol{G} H \boldsymbol{G}^\top u > \gamma^2 \left(\sqrt{\mathrm{tr}(H)} + \sqrt{2\|H\|_{\mathrm{op}} t}\right)^2\right\} \leq e^{-t}.$$

Since $|\mathbb{N}| \leq 5^n$, see [90, Lemma 5.2], choosing $t = 3n$ and taking a union bound over $u \in \mathbb{N}$ completes the proof. ■

Another useful concentration inequality of the operator norm of the random matrices with i.i.d. sub-Gaussian rows is stated in the following lemma. This is an immediate result of [90, Remark 5.40].

Lemma 48. *Let $\boldsymbol{G}$ be n by d matrix whose rows are i.i.d. γ sub-Gaussian random vectors with covariance matrix Σ_Y. Then for every $t \geq 0$, with probability at least $1 - 2e^{-ct^2}$,*

$$\left\|\frac{1}{n}\boldsymbol{G}^\top \boldsymbol{G} - \Sigma_Y\right\|_{\mathrm{op}} \leq \max\left\{\delta, \delta^2\right\} \|\Sigma_Y\|_{\mathrm{op}},$$

with $\delta = C\sqrt{d/n} + t/\sqrt{n}$ where $c = c(\gamma)$ and $C = C(\gamma)$ are positive constants depending on γ.

The deviation inequalities of the inner product of two random vectors with independent sub-Gaussian elements are well-known; we state the one in [22] for completeness.

Lemma 49. *[22, Lemma 10] Let $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$ be any two sequences, each with zero mean independent γ_x sub-Gaussian and γ_y sub-Gaussian elements. Then, for some absolute constant $c > 0$, we have*

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{t=1}^n (X_t Y_t - \mathbb{E}[X_t Y_t])\right| \leq \gamma_x \gamma_y t\right\} \geq 1 - 2\exp\left\{-c\min\left(t^2, t\right)n\right\}.$$

In particular, when $\log p \leq n$, one has

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{t=1}^n (X_t Y_t - \mathbb{E}[X_t Y_t])\right| \leq C\sqrt{\frac{\log(p \vee n)}{n}}\right\} \geq 1 - 2(p \vee n)^{-c}$$

where $c \geq 2$ and $C = C(\gamma_x, \gamma_y, c)$ are some positive constants.

B.4 The LOVE Algorithm

For the reader's convenience, we give the specifics of estimating $\widehat{A}$ in the Essential Regression model, as developed in [26]. The first step is estimation of the number of latent factors, K, and the partition of pure variables, $\mathcal{I}$, which is achieved by Algorithm B.4 below.

[ht] [1] PureVar$\widehat{\Sigma}$, δ $\widehat{\mathcal{I}} \leftarrow \varnothing$. $i \in [p]$ $\widehat{I^{(i)}} \leftarrow \{l \in [p] \setminus \{i\} : \max_{j \in [p]\setminus\{i\}} |\widehat{\Sigma}_{ij}| \leq |\widehat{\Sigma}_{il}| + 2\delta\}$ $Pure(i) \leftarrow True$. $j \in \widehat{I^{(i)}}$ $\left||\widehat{\Sigma}_{ij}| - \max_{k \in [p]\setminus\{j\}} |\widehat{\Sigma}_{jk}|\right| > 2\delta$ $Pure(i) \leftarrow False$, **break** $Pure(i)$ $\widehat{I^{(i)}} \leftarrow \widehat{I^{(i)}} \cup \{i\}$ $\widehat{\mathcal{I}} \leftarrow$ MERGE$(\widehat{I^{(i)}}, \widehat{\mathcal{I}})$ $\widehat{\mathcal{I}}$ and $\widehat{K}$ as the number of sets in $\widehat{\mathcal{I}}$ Merge$\widehat{I^{(i)}}$, $\widehat{\mathcal{I}}$ $G \in \widehat{\mathcal{I}}$ $\widehat{\mathcal{I}}$ is a collection of sets $G \cap \widehat{I^{(i)}} \neq \varnothing$ $G \leftarrow G \cap \widehat{I^{(i)}}$ Replace $G \in \widehat{\mathcal{I}}$ by $G \cap \widehat{I^{(i)}}$ $\widehat{\mathcal{I}}$ $\widehat{I^{(i)}} \in \widehat{\mathcal{I}}$ add $\widehat{I^{(i)}}$ in $\widehat{\mathcal{I}}$ $\widehat{\mathcal{I}}$ Given estimates $\widehat{K}$ and $\widehat{\mathcal{I}}$ as outputs of Algorithm 1, we compute, for each $a \in [\widehat{K}]$ and $b \in [\widehat{K}] \setminus \{a\}$,

$$\left[\widehat{\Sigma}_Z\right]_{aa} = \frac{1}{|\widehat{I}_a|(|\widehat{I}_a| - 1)} \sum_{i,j \in \widehat{I}_a, i \neq j} |\widehat{\Sigma}_{ij}|, \quad \left[\widehat{\Sigma}_Z\right]_{ab} = \frac{1}{|\widehat{I}_a||\widehat{I}_b|} \sum_{i \in \widehat{I}_a, j \in \widehat{I}_b} \widehat{A}_{ia}\widehat{A}_{ib}\widehat{\Sigma}_{ij}, \tag{B.32}$$

to form the estimator $\widehat{\Sigma}_Z$ of Σ_Z.

The submatrix $\widehat{A}_{\widehat{I}}$ is then constructed as follows. For each $k \in [\widehat{K}]$ and the estimated pure variable set $\widehat{I}_k$,

$$\text{Pick an element } i \in \widehat{I}_k \text{ at random, and set } \widehat{A}_{i\cdot} = e_k; \tag{B.33}$$

$$\text{For the remaining } j \in \widehat{I}_k \setminus \{i\}, \text{ set } \widehat{A}_{j\cdot} = \text{sign}(\widehat{\Sigma}_{ij}) \cdot e_k. \tag{B.34}$$

Letting $\widehat{J} = [p] \setminus \widehat{I}$, to construct the remaining submatrix $\widehat{A}_{\widehat{J}}$, we use the Dantzig-type estimator $\widehat{A}_D$ proposed in [26] given by

$$\widehat{A}_{j\cdot} = \arg\min_{\beta^j} \left\{ \|\beta^j\|_1 : \left\| \widehat{\Sigma}_Z \beta^j - (\widehat{A}_{\widehat{I}}^\top \widehat{A}_{\widehat{I}})^{-1} \widehat{A}_{\widehat{I}}^\top \widehat{\Sigma}_{\widehat{I}j} \right\|_\infty \leq \mu \right\} \tag{B.35}$$

for any $j \in \widehat{J}$, with tuning parameter $\mu = O(\sqrt{\log(p \vee n)/n})$. The estimator $\widehat{A}$ enjoys the optimal convergence rate of $\max_{j\in[p]} \|\widehat{A}_{j\cdot} - A_{j\cdot}\|_q$ for any $1 \le q \le \infty$ [26, Theorem 5].

B.5 More Existing Literature on Factor Models

We discuss in this section some related work on factor models which might be used to establish results of the excess risk of PCR.

By treating X and Y jointly from model 2.1 as an augmented factor model

$$\widetilde{X} := \begin{bmatrix} Y \\ X \end{bmatrix} = \begin{bmatrix} \beta^\top \\ A \end{bmatrix} Z + \begin{bmatrix} \varepsilon \\ W \end{bmatrix},$$

the fit $\widehat{\boldsymbol{Y}}$ is constructed by regressing $\boldsymbol{Y}$ onto $\widetilde{\boldsymbol{X}}\widetilde{\mathbf{U}}_K$ where $\widetilde{\mathbf{U}}_K$ is the matrix of the first K right singular vectors of $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_{1\bullet}^\top, \ldots, \widetilde{\boldsymbol{X}}_{n\bullet}^\top)^\top$. [8] shows that

$$V_t^{-1/2}\left(\widehat{\boldsymbol{Y}}_t - \boldsymbol{Z}_{t\bullet}^\top \beta\right) \to N(0,1), \qquad \text{for any } 1 \le t \le n \tag{B.36}$$

for a variance term V_t. The uniform convergence rate of $\widehat{\boldsymbol{Y}}_t - \boldsymbol{Z}_{t\bullet}^\top \beta$ over $1 \le t \le n$ is further derived in [42]. These element-wise results for *in-sample* prediction could, in principle, be extended to out-of-sample prediction, via additional arguments, but is not treated in the aforementioned works.

We now comment on the main differences between our Corollary 19 and the aforementioned results. The existing results are all established under conditions including $K = O(1)$, $\|\beta\|_2^2 = O(1)$, $p \to \infty$, and (2.29), The uniform consistency in [42] additionally requires $n = o(p^2)$. As a result, all previous results are asymptotic statements as $n, p \to \infty$.

By contrast, our Corollaries 18, 19 and 21 are non-asymptotic statements which hold for any finite K, n and p. Moreover, they only requires the sub-Gaussian tail assumptions in Definition 2.2.1 and $K \log n \lesssim n$. As detailed in Section 2.3.2, our conditions on the signal $\lambda_K(A\Sigma_Z A^\top)$ are much weaker than (2.29) to derive the risk of PCR-K.

Under condition (2.29), as assumed in the aforementioned literature, the prediction risk in our Corollary 19 reduces to

$$\mathbb{R}(\mathbf{U}_K) - \sigma^2 = O_p\left(\frac{\sigma^2}{n} + \frac{\|\Sigma_W\|_{\text{op}}}{p} + \frac{\|\Sigma_W\|_{\text{op}}}{n}\right).$$

This rate coincides with that of V_t, introduced in (B.36). Under conditions in [42], their results (see, for instance, Corollary 3.1) imply

$$\max_{1\le t\le n}\left|\widehat{\boldsymbol{Y}}_t - \boldsymbol{Z}_{t\bullet}^\top \beta\right|^2 = O_p\left((\log n)^{2/r_2}\,\frac{\log p}{n} + \frac{n^{1/2}}{p}\right)$$

for some constant $r_2 > 0$, which is slower than our rate.

APPENDIX C

APPENDIX OF CHAPTER 3

C.1 Proof of Theorem 26

We work on the event where $\widehat{K} = K$ and $\widehat{K}' = K'$. We assume the minimizer P in (3.15) and (3.17) is the identity I_K without loss of generality. Define

$$\widetilde{\boldsymbol{\theta}} = \sum_{a=1}^{K} \theta_a \delta_{\widehat{\boldsymbol{\tau}}^{(a)}},$$

and similarly,

$$\widetilde{\boldsymbol{\theta}}' = \sum_{a=1}^{K} \theta'_a \delta_{\widehat{\boldsymbol{\tau}}'^{(a)}}.$$

Note that $\widehat{d}^{\text{mix}}$ and d^{mix} are both equal to the Total Variation distance on Δ_p, so we only use d^{mix} in the proof below. Furthermore, on the event where $\widehat{K} = K$ and $\widehat{K}' = K'$, both d^{cluster} and $\widehat{d}^{\text{cluster}}$ are both equal to the Wasserstein distance $W_1(\mu, \nu; d^{\text{mix}})$ defined for $\mu, \nu \in \mathcal{D}_{K \vee K', p}$, therefore for notational simplicity we only use the notation d^{cluster} in the proof below. By the triangle inequality for the Wasserstein distance,

$$\begin{aligned} W_1(\boldsymbol{\theta}, \boldsymbol{\theta}', d^{\text{cluster}}) \leq\ & W_1(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}; d^{\text{cluster}}) + W_1(\widetilde{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}; d^{\text{cluster}}) + W_1(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}'; d^{\text{cluster}}) \\ & + W_1(\widehat{\boldsymbol{\theta}}', \widetilde{\boldsymbol{\theta}}'; d^{\text{cluster}}) + W_1(\widetilde{\boldsymbol{\theta}}', \boldsymbol{\theta}'; d^{\text{cluster}}). \end{aligned}$$

Using the triangle inequality again,

$$\begin{aligned} W_1(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\theta}}', d^{\text{cluster}}) \leq\ & W_1(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}; d^{\text{cluster}}) + W_1(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}; d^{\text{cluster}}) + W_1(\boldsymbol{\theta}, \boldsymbol{\theta}'; d^{\text{cluster}}) \\ & + W_1(\boldsymbol{\theta}', \widetilde{\boldsymbol{\theta}}'; d^{\text{cluster}}) + W_1(\widetilde{\boldsymbol{\theta}}', \widehat{\boldsymbol{\theta}}'; d^{\text{cluster}}). \end{aligned}$$

Combining the previous two displays and using the upper bound of the Wasserstein distance by the Total Variation distance (see [47], for example), we

find

$$\begin{aligned}|W_1(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\theta}}';d^{\text{cluster}}) - W_1(\boldsymbol{\theta},\boldsymbol{\theta}';d^{\text{cluster}})| &\le W_1(\widehat{\boldsymbol{\theta}},\widetilde{\boldsymbol{\theta}};d^{\text{cluster}}) + W_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta};d^{\text{cluster}})+\\ &\quad + W_1(\boldsymbol{\theta}',\widetilde{\boldsymbol{\theta}}';d^{\text{cluster}}) + W_1(\widetilde{\boldsymbol{\theta}}',\widehat{\boldsymbol{\theta}}';d^{\text{cluster}})\\ &\le \text{TV}(\widehat{\boldsymbol{\theta}},\widetilde{\boldsymbol{\theta}}) + W_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta};d^{\text{cluster}})+\\ &\quad + W_1(\boldsymbol{\theta}',\widetilde{\boldsymbol{\theta}}';d^{\text{cluster}}) + \text{TV}(\widetilde{\boldsymbol{\theta}}',\widehat{\boldsymbol{\theta}}').\end{aligned}$$

We bound these four terms in (C.7), (C.8), (C.9), and (C.10) below, which leads to

$$\begin{aligned}|W_1(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\theta}}';d^{\text{cluster}}) - W_1(\boldsymbol{\theta},\boldsymbol{\theta}';d^{\text{cluster}})| &\lesssim \max_{k\in[K]}\|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1 + \max_{k\in[K']}\|\widehat{A}'_{\cdot k} - A'_{\cdot k}\|_1\\ &\quad + \theta_{\min}^{-1}\max_{i\in[n]}\|\widehat{T}^{(i)} - T^{(i)}\|_1 + \theta_{\min}'^{-1}\max_{i\in[n']}\|\widehat{T}'^{(i)} - T'^{(i)}\|_1.\end{aligned}$$

Combining this with the rates (3.15) and (3.17) gives the final result.

Bound of $W_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta};d^{\text{cluster}})$

By definition,

$$W_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta};d^{\text{cluster}}) = \inf_{w\in\Sigma_W(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})}\sum_{a,b=1}^{K} w(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(b)})\cdot W_1(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(b)};d^{\text{mix}}).$$

Define

$$w^* = \sum_{a=1}^{K}\theta_a\delta_{\widehat{\boldsymbol{T}}^{(a)}}\times\delta_{\boldsymbol{T}^{(a)}}, \tag{C.1}$$

and note that $w^*\in\Sigma_W(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta})$. Thus,

$$\begin{aligned}W_1(\widetilde{\boldsymbol{\theta}},\boldsymbol{\theta};d^{\text{cluster}}) &\le \sum_{a,b=1}^{K} w^*(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(b)})\cdot W_1(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(a)};d^{\text{mix}})\\ &= \sum_{a=1}^{K}\theta_a\cdot W_1(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(a)};d^{\text{mix}})\\ &\le \max_{a\in[K]} W_1(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(a)};d^{\text{mix}})\cdot\sum_{a=1}^{K}\theta_a\\ &= \max_{a\in[K]} W_1(\widehat{\boldsymbol{T}}^{(a)},\boldsymbol{T}^{(a)};d^{\text{mix}})\end{aligned} \tag{C.2}$$

Defining

$$\widetilde{\boldsymbol{\mathcal{T}}}^{(a)} := \sum_{k=1}^{K} \boldsymbol{\mathcal{T}}^{(k)} \delta_{\widehat{A}_{\cdot k}}, \tag{C.3}$$

the triangle inequality for the Wasserstein distance gives

$$W_1(\widehat{\widetilde{\boldsymbol{\mathcal{T}}}}^{(a)}, \boldsymbol{\mathcal{T}}^{(a)}; d^{\text{mix}}) \le W_1(\widehat{\widetilde{\boldsymbol{\mathcal{T}}}}^{(a)}, \widetilde{\boldsymbol{\mathcal{T}}}^{(a)}; d^{\text{mix}}) + W_1(\widetilde{\boldsymbol{\mathcal{T}}}^{(a)}, \boldsymbol{\mathcal{T}}^{(a)}; d^{\text{mix}}). \tag{C.4}$$

For the rightmost term, we have

$$W_1(\widetilde{\boldsymbol{\mathcal{T}}}^{(a)}, \boldsymbol{\mathcal{T}}^{(a)}; d^{\text{mix}}) = \inf_{w \in \Sigma_W(\widetilde{\boldsymbol{\mathcal{T}}}^{(a)}, \boldsymbol{\mathcal{T}}^{(a)})} \sum_{k,l=1}^{K} w(\widehat{A}_{\cdot k}, A_{\cdot l}) \cdot d^{\text{mix}}(\widehat{A}_{\cdot k}, A_{\cdot l}).$$

Using the coupling

$$\sum_{k=1}^{K} \boldsymbol{\mathcal{T}}_k^{(a)} \cdot \delta_{\widehat{A}_{\cdot k}} \times \delta_{A_{\cdot k}} \in \Sigma_W(\widetilde{\boldsymbol{\mathcal{T}}}^{(a)}, \boldsymbol{\mathcal{T}}^{(a)}),$$

we have

$$\begin{aligned} W_1(\widetilde{\boldsymbol{\mathcal{T}}}^{(a)}, \boldsymbol{\mathcal{T}}^{(a)}; d^{\text{mix}}) &\le \sum_{k=1}^{K} \boldsymbol{\mathcal{T}}_k^{(a)} \cdot d^{\text{mix}}(\widehat{A}_{\cdot k}, A_{\cdot k}) \\ &\le \max_{k \in [K]} \frac{1}{2} \|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1 \end{aligned} \tag{C.5}$$

For the first term in the right hand side of (C.4),

$$\begin{aligned} W_1(\widehat{\widetilde{\boldsymbol{\mathcal{T}}}}^{(a)}, \widetilde{\boldsymbol{\mathcal{T}}}^{(a)}; d^{\text{mix}}) &\le \text{TV}(\widehat{\widetilde{\boldsymbol{\mathcal{T}}}}^{(a)}, \widetilde{\boldsymbol{\mathcal{T}}}^{(a)}) \\ &= \frac{1}{2} \sum_{k=1}^{K} |\widehat{\boldsymbol{\mathcal{T}}}_k^{(a)} - \boldsymbol{\mathcal{T}}_k^{(a)}| \\ &= \frac{1}{2} \sum_{k=1}^{K} \left| \sum_{i=1}^{n} (\widehat{\gamma}_i^{(a)} \widehat{T}_k^{(i)} - \gamma_i^{(a)} T_k^{(i)}) \right| \\ &\le \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} |T_k^{(i)}| |\widehat{\gamma}_i^{(a)} - \gamma_i^{(a)}| + \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} |\widehat{\gamma}_i^{(a)}| |\widehat{T}_k^{(i)} - T_k^{(i)}| \\ &\le \frac{1}{2} \|\widehat{\gamma}^{(a)} - \gamma^{(a)}\|_1 + \frac{1}{2} \max_{i \in [n]} \|\widehat{T}^{(i)} - T^{(i)}\|_1, \end{aligned} \tag{C.6}$$

where in the second line we use that $\widehat{\widetilde{\boldsymbol{\mathcal{T}}}}^{(a)}$ and $\widetilde{\boldsymbol{\mathcal{T}}}^{(a)}$ are both discrete distributions on $(\widehat{A}_{\cdot k})_{k \in [K]}$ with weights $(\widehat{\boldsymbol{\mathcal{T}}}_k^{(a)})_{k \in [K]}$ and $(\boldsymbol{\mathcal{T}}_k^{(a)})_{k \in [K]}$, respectively, and in the final

step we used $\|\widehat{\gamma}^{(a)}\|_1 = \|T^{(i)}\|_1 = 1$. Note that using the definition of $\gamma^{(a)}$ and θ_a, we can write

$$\gamma_i^{(a)} = \frac{T_a^{(i)}}{\sum_{j=1}^n T_a^{(j)}} = \frac{1}{n}\frac{T_a^{(i)}}{\theta_a}.$$

Similarly,

$$\widehat{\gamma}_i^{(a)} = \frac{1}{n}\frac{\widehat{T}_a^{(i)}}{\widehat{\theta}_a}.$$

Using this, we find for any $i \in [n]$,

$$\begin{aligned}
|\widehat{\gamma}_i^{(a)} - \gamma_i^{(a)}| &= \frac{1}{n}\left|\frac{\widehat{T}_a^{(i)}}{\widehat{\theta}_a} - \frac{T_a^{(i)}}{\theta_a}\right| \\
&= \frac{1}{n}\left|\frac{\widehat{T}_a^{(i)}}{\widehat{\theta}_a} - \frac{\widehat{T}_a^{(i)}}{\theta_a} + \frac{\widehat{T}_a^{(i)}}{\theta_a} - \frac{T_a^{(i)}}{\theta_a}\right| \\
&\le \frac{1}{n}\frac{|\widehat{T}_a^{(i)} - T_a^{(i)}|}{\theta_a} + \frac{1}{n}\frac{\widehat{T}_a^{(i)}}{\widehat{\theta}_a}\frac{|\widehat{\theta}_a - \theta_a|}{\theta_a}.
\end{aligned}$$

Summing over $i \in [n]$ and using $n^{-1}\sum_i \widehat{T}_a^{(i)} = \widehat{\theta}_a$, we get

$$\begin{aligned}
\|\widehat{\gamma}^{(a)} - \gamma^{(a)}\|_1 &\le \frac{1}{n}\sum_{i=1}^n \frac{|\widehat{T}_a^{(i)} - T_a^{(i)}|}{\theta_a} + \frac{|\widehat{\theta}_a - \theta_a|}{\theta_a} \\
&= \frac{1}{n}\sum_{i=1}^n \frac{|\widehat{T}_a^{(i)} - T_a^{(i)}|}{\theta_a} + \frac{1}{\theta_a}\left|\frac{1}{n}\sum_{i=1}^n (\widehat{T}_a^{(i)} - T_a^{(i)})\right| \\
&\le \frac{2}{n}\sum_{i=1}^n \frac{|\widehat{T}_a^{(i)} - T_a^{(i)}|}{\theta_a} \\
&\le 2\max_{i\in[n]} \frac{|\widehat{T}_a^{(i)} - T_a^{(i)}|}{\theta_a}
\end{aligned}$$

Then,

$$\max_{a\in[K]} \|\widehat{\gamma}^{(a)} - \gamma^{(a)}\|_1 \le 2\max_{i\in[n]}\max_{a\in[K]} \frac{|\widehat{T}_a^{(i)} - T_a^{(i)}|}{\theta_a} \le \frac{2}{\theta_{\min}}\max_{i\in[n]} \|\widehat{T}^{(i)} - T^{(i)}\|_1.$$

Combining this with (C.6) we find

$$\max_{a\in[K]} W_1(\widehat{\boldsymbol{\mathcal{T}}}^{(a)}, \widetilde{\boldsymbol{\mathcal{T}}}^{(a)}; d^{\text{mix}}) \le \left(\frac{1}{2} + \frac{1}{\theta_{\min}}\right)\max_{i\in[n]} \|\widehat{T}^{(i)} - T^{(i)}\|_1$$

Combining this with (C.2), (C.4), and (C.5), we find

$$W_1(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}; d^{\text{cluster}}) \le \left(\frac{1}{2} + \frac{1}{\theta_{\min}}\right)\max_{i\in[n]} \|\widehat{T}^{(i)} - T^{(i)}\|_1 + \max_{k\in[K]} \frac{1}{2}\|\widehat{A}_{\cdot k} - A_{\cdot k}\|_1 \tag{C.7}$$

An analogous proof for the second corpus gives

$$W_1(\widetilde{\boldsymbol{\theta}'}, \boldsymbol{\theta}'; d^{\text{cluster}}) \leq \left(\frac{1}{2} + \frac{1}{\theta'_{\min}}\right) \max_{i \in [n']} \|\widehat{T}'^{(i)} - T'^{(i)}\|_1 + \max_{k \in [K]} \frac{1}{2} \|\widehat{A}'_{\cdot k} - A'_{\cdot k}\|_1. \tag{C.8}$$

Bounding TV($\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}$) and TV($\widehat{\boldsymbol{\theta}'}, \widetilde{\boldsymbol{\theta}'}$)

Using the fact that $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ are both discrete distributions on $(\widehat{\boldsymbol{\mathcal{T}}}^{(a)})_{a \in [K]}$ with weights $(\widehat{\theta}_a)_{a \in [K]}$ and $(\theta_a)_{a \in [K]}$, respectively,

$$\begin{aligned}
\text{TV}(\widehat{\boldsymbol{\theta}}, \widetilde{\boldsymbol{\theta}}) &= \sum_{k=1}^{K} |\widehat{\theta}_k - \theta_k| = \frac{1}{n} \sum_{k=1}^{K} \left| \sum_{i=1}^{n} (\widehat{T}_k^{(i)} - T_k^{(i)}) \right| \\
&\leq \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} |\widehat{T}_k^{(i)} - T_k^{(i)}| \\
&= \frac{1}{n} \sum_{i=1}^{n} \|\widehat{T}^{(i)} - T^{(i)}\|_1 \\
&\leq \max_{i \in [n]} \|\widehat{T}^{(i)} - T^{(i)}\|_1.
\end{aligned} \tag{C.9}$$

Following the same approach for the second corpus, we find

$$\text{TV}(\widehat{\boldsymbol{\theta}'}, \widetilde{\boldsymbol{\theta}'}) \leq \max_{i \in [n']} \|\widehat{T}'^{(i)} - T'^{(i)}\|_1. \tag{C.10}$$

■

BIBLIOGRAPHY

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

[2] Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

[3] Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-kai Liu. A spectral algorithm for latent dirichlet allocation. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 917–925. Curran Associates, Inc., 2012.

[4] T. 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, Berkeley, Calif., 1956. University of California Press.

[5] Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, pages 280–288, 2013.

[6] Sanjeev Arora, Rong Ge, Frederic Koehler, Tengyu Ma, and Ankur Moitra. Provable algorithms for inference in topic models. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2859–2867, New York, New York, USA, 20–22 Jun 2016. PMLR.

[7] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012, IEEE 53rd Annual Symposium*, pages 1–10. IEEE, 2012.

[8] Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003.

[9] Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[10] Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006.

[11] Jushan Bai and Serena Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317, 2008. Honoring the research contributions of Charles R. Nelson.

[12] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

[13] 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.

[14] 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, 2019.

[15] Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *In arXiv:1806.05161*, 2018.

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

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

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

[19] Anil Bhattacharya and David B. Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

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

[21] Xin Bing, Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp.

Likelihood estimation of sparse topic distributions in topic models and its applications to wasserstein document distance calculations, 2021.

[22] Xin Bing, Florentina Bunea, and Marten Wegkamp. Inference in interpretable latent factor regression models. *In arXiv:1905.12696*, 2019.

[23] Xin Bing, Florentina Bunea, and Marten Wegkamp. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli*, 26(3):1765–1796, 08 2020.

[24] Xin Bing, Florentina Bunea, and Marten Wegkamp. Optimal estimation of sparse topic models. *Journal of Machine Learning Research*, 21(177):1–45, 2020.

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

[26] Xin Bing, Florentina Bunea, Ning Yang, and Marten Wegkamp. Adaptive estimation in structured factor models with applications to overlapping clustering. *To appear in the Annals of Statistics*, 2020.

[27] Xin Bing and Marten H. Wegkamp. Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.*, 47(6):3157–3184, 12 2019.

[28] Victor Bittorf, Benjamin Recht, Christopher Re, and Joel A Tropp. Factoring nonnegative matrices with linear programs. *arXiv:1206.1270*, 2012.

[29] David M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 55:77–84, 2012.

[30] 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.

[31] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[32] 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.

[33] Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Interpolation under latent factor regression models. *In arXiv:2002.02525*, 2020.

[34] Florentina Bunea and Luo Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli*, 21(2):1200–1230, 05 2015.

[35] 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.

[36] Michael H. Coen, M. Hidayath Ansari, and Nathanael Fillmore. Comparing clusterings in space. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 231–238, Madison, WI, USA, 2010. Omnipress.

[37] Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation (to appear). *Journal of the American Statistical Association*, 2019.

[38] W. Ding, P. Ishwar, and V. Saligrama. Most large topic models are approximately separable. In *2015 Information Theory and Applications Workshop (ITA)*, pages 199–203, 2015.

[39] Weicong Ding, Mohammad Hossein Rohban, Prakash Ishwar, and Venkatesh Saligrama. Topic discovery through data dependent and random projections. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1202–1210, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[40] Jianqing Fan, Yuan Liao, and Martina Mincheva. High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, 39(6):3320–3356, 12 2011.

[41] 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, 2013.

[42] 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, 2013.

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

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

[45] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *AISTATS*, 2019.

[46] Richard Fothergill, Paul Cook, and Timothy Baldwin. Evaluating a topic modelling approach to measuring corpus similarity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 273–279, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[47] Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[48] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[49] 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.

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

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

[52] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1501–1509. PMLR, 06–11 Aug 2017.

[53] Harold Hotelling. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79, 1957.

[54] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, June 2014.

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

[56] 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.

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

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

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

[60] 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.

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

[62] Tracy Zheng Ke and Minzhe Wang. A new svd approach to optimal topic estimation. *arXiv:1704.07016*, 2017.

[63] Bryan Kelly and Seth Pruitt. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294 – 316, 2015. High Dimensional Problems in Econometrics.

[64] Maurice G. Kendall. *A course in multivariate analysis*. Hafner Pub. Co., 1957.

[65] Adam Kilgarriff. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133, 2001.

[66] Jon Kleinberg and Mark Sandler. Using mixture models for collaborative filtering. *Journal of Computer and System Sciences*, 74(1):49–69, 2008.

[67] Olga Klopp, Maxim Panov, Suzanne Sigalla, and Alexandre Tsybakov. Assigning topics to documents by successive projections, 2021.

[68] Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.*, 40(2):694–726, 04 2012.

[69] 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.

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

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

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

[73] 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.

[74] Johan Segers Marc Hallin, Gilles Mordant. Multivariate goodness-of-fit tests based on wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328–1371, 2021.

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

[76] 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.

[77] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *In arXiv:1911.01544*, 2019.

[78] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overpa-

rameterized regimes: Does the loss function matter? *In arXiv:2005.08054*, 2020.

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

[80] Xuanlong Nguyen. Borrowing strength in hierarchical bayes: Posterior concentration of the dirichlet base measure. *Bernoulli*, 22(3):1535–1571, 2016.

[81] Victor M. Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer Nature, 2020.

[82] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. *English Gigaword Fifth Edition LDC2011T07*. Linguistic Data Consortium, Philadelphia, 2011.

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

[84] D. Pollard. Quantization and the method ofk-means. *IEEE Transactions on Information Theory*, 28(2):199–205, 1982.

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

[86] 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, 2002.

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

[88] 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.

[89] Carla Tameling, Max Sommerfeld, and Axel Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29, 07 2017.

[90] Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012.

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

[92] Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.

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

[94] Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.