Journal of Machine Learning Research 22 (2021) 1-50 Submitted 7/20; Revised 4/21; Published 7/21

Prediction Under Latent Factor Regression: Adaptive PCR, Interpolating Predictors and Beyond

Xin Bing XB43@CORNELL.EDU
Florentina Bunea FB238@CORNELL.EDU
Seth Strimas-Mackey SCS324@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14850, USA

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

Editor: Arnak Dalalyan

Abstract

This work is devoted to the finite sample prediction risk analysis of a class of linear predictors of a response $Y \in \mathbb{R}$ from a high-dimensional random vector $X \in \mathbb{R}^p$ when (X, Y) follows a latent factor regression model generated by a unobservable latent vector Z of dimension less than p. 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 known risk bounds for the minimum-norm interpolating predictor, which has received renewed attention in the past two years, 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. We conclude with a detailed simulation study to support and complement our theoretical results.

Keywords: High-dimensional regression, latent factor model, principal component regression, interpolating predictor, model selection

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

$$\begin{aligned} Y &= Z^\top \beta + \varepsilon, \\ X &= AZ + W, \end{aligned} \tag{1}$$

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 (Kendall, 1957; Hotelling, 1957). 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

$$\begin{aligned} \widehat{Y}^*_{\boldsymbol{U}_k} &= X_*^\top \boldsymbol{U}_k \left[\boldsymbol{U}_k^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{U}_k\right]^+ \boldsymbol{U}_k^\top \boldsymbol{X}^\top \boldsymbol{Y} \\ &= X_*^\top \boldsymbol{U}_k \left[\boldsymbol{X}\boldsymbol{U}_k\right]^+ \boldsymbol{Y}, \end{aligned} \tag{2}$$

where $\boldsymbol{U}_k$ 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 (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 $p, n \to \infty$ (Stock and Watson, 2002; Bai and Ng, 2006), and finite sample / asymptotic risk bounds on the in-sample prediction accuracy of PCR-K (Bai, 2003; Bair et al., 2006; Fan et al., 2013; Kelly and Pruitt, 2015; Fan et al., 2017) in identifiable factor models with known and fixed K.

To the best of our knowledge, finite sample prediction risk bounds for $\widehat{Y}^*_{\boldsymbol{U}_k}$, 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 3, that establishes a finite sample prediction risk bound for linear predictors of the general form

$$\widehat{Y}^*_{\widehat{B}} = X_*^\top \widehat{B} \left(\widehat{B}^\top \boldsymbol{X}^\top \boldsymbol{X} \widehat{B}\right)^+ \widehat{B}^\top \boldsymbol{X}^\top \boldsymbol{Y}, \tag{3}$$

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 $\widehat{B} = \boldsymbol{U}_k$, but of offering a unifying analysis of other prediction schemes of the form (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

(Montanari et al., 2019; Bunea et al., 2020; Muthukumar et al., 2019, 2020; Hastie et al., 2019; Feldman, 2019; Belkin et al., 2019a,b, 2018a,b,c; Bartlett et al., 2019; Liang and Rakhlin, 2018). 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 3 can be used to analyze predictors directly tailored to model (1), which are shown formally to be of type (3) in Section 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 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.

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 $\mathcal{R}(\widehat{B})$, defined as

$$\mathcal{R}(\widehat{B}) \coloneqq \mathbb{E}[(Y_* - \widehat{Y}^*_{\widehat{B}})^2], \tag{4}$$

provably approaches an optimal risk benchmark, as n and p grow, with particular attention given to the case $p > n$. The expectation in (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, we state the risk benchmark in Lemma 2 and prove a master theorem, and our main theoretical result, Theorem 3. It provides a finite sample bound on $\mathcal{R}(\widehat{B})$, for generic $\widehat{B}$, when (X, Y) follow a factor regression model (1) that is fully introduced in Section 2.1.

The risk bound (14) of Theorem 3 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 15 in Appendix B 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 3, as a general result, is in reducing the difficult task of bounding $\mathcal{R}(\widehat{B})$ 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 3 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 6 provides the rate

$$\mathcal{R}(\boldsymbol{U}_{\widehat{s}}) - \sigma^2 \lesssim (K + \log n)\frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1}\beta. \tag{5}$$

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 2 of Section 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 3.1, which we show in Corollary 9 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 4.1 we use the master theorem to recover risk bounds for the Generalized Least Squares predictor (GLS), independently derived in Bunea et al. (2020). 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 4.2 we consider a particular identifiable factor regression model, the Essential Regression model introduced in Bing et al. (2019). The identifiability assumptions employ a type of errors-in-variables parametrization of A, described in Section 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, $1 \le k \le K$ (Bing et al., 2020). Thus, of independent interest, prediction in Essential Regression is prediction via latent cluster centers. We show formally in Section 4.2 that this model specification leads to predictors of type (3), with $\widehat{B} = \widehat{A}$, for an appropriate estimator $\widehat{A}$ of A. We provide a finite sample prediction bound in Theorem 12, as an application of Theorem 3. 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 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 3—4 is given in Table 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

$r_e(\Sigma_W) \coloneqq \mathrm{tr}(\Sigma_W)/\|\Sigma_W\|_{\mathrm{op}}$ is the reduced effective rank of Σ_W, the covariance matrix of W from model (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 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 4.3.

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

Table 1: Summary of bounds on $\mathcal{R}(\widehat{B}) - \sigma^2$, where $\mathcal{R}(\widehat{B})$ is defined in (4), for Principal Component Regression (PCR), Generalized Least Squares (GLS), and Essential Regression (ER), stated under simplifying assumptions described in Section 4.3. The second column gives the choice of $\widehat{B}$ corresponding to each method. All three bounds follow from the main Theorem 3.

We conclude the paper with Section 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 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 $\lambda_1(Q) \geq \lambda_2(Q) \geq \cdots \geq \lambda_p(Q)$ 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 $a_n \leq C b_n$. 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 $m \geq 1$, we let $[m] = \{1, 2, \ldots, m\}$. Lastly, we use c, c', C, C' to denote positive and finite absolute constants that unless otherwise indicated can change from line to line.

2. Bounding the Risk $\mathcal{R}(\widehat{B})$

In this section we derive and discuss bounds on the risk $\mathcal{R}(\widehat{B})$ defined in (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.

2.1 Preliminaries

As the risk $\mathcal{R}(\widehat{B})$ 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 (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 (1) with respect to θ.

Definition 1 ((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 (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 $\mathbb{E}_\theta[\widetilde{W}\widetilde{W}^\top] = \boldsymbol{I}_p$.

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

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 $\mathcal{R}(\widehat{B}) \coloneqq \mathbb{E}_\theta[(Y_* - \widehat{Y}^*_{\widehat{B}})^2]$ 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 1. We note that $\mathcal{R}(\widehat{B})$ depends on θ, but we suppress this dependence in the notation for simplicity.

2.2 Benchmark of $\mathcal{R}(\widehat{B})$

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

$$\alpha^* := \arg\min_{\alpha} \mathbb{E}\left[(Y_* - X_*^\top \alpha)^2\right] = [\mathrm{Cov}(X)]^+ \mathrm{Cov}(X, Y) \tag{6}$$

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

$$\begin{aligned} \mathcal{R}(\widehat{B}) &= \mathbb{E}_\theta\left[(Y_* - X_*^\top \alpha^*)^2\right] + \mathbb{E}_\theta\left[(X_*^\top \alpha^* - \widehat{Y}^*_{\widehat{B}})^2\right] \\ &= \sigma^2 + \mathbb{E}_\theta\left[(Z_*^\top \beta - X_*^\top \alpha^*)^2\right] + \mathbb{E}_\theta\left[(X_*^\top \alpha^* - \widehat{Y}^*_{\widehat{B}})^2\right] \\ &= \sigma^2 + \mathbb{E}_\theta\left[(Z_*^\top \beta - \widehat{Y}^*_{\widehat{B}})^2\right]. \end{aligned} \tag{7}$$

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 $\mathcal{R}(\widehat{B}) - \sigma^2$ 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 (7) is the minimal risk incurred by predicting $Z_*^\top \beta$ by $X_*^\top \alpha^*$, with an observable X_*. Display (7) shows that it is a population level cost that is incurred in any risk analysis of a predictor of type (3) performed under FRM(θ). Lemma 2 below quantifies its size, and makes use of the signal-to-noise ratio given by

$$\xi \coloneqq \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_W\|_{\mathrm{op}}. \tag{8}$$

Its proof can be found in Appendix B.1.

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

$$\frac{\xi}{1+\xi}\beta^\top (A^\top \Sigma_W^{-1} A)^{-1}\beta \le \mathbb{E}_\theta\left[(Z_*^\top \beta - X_*^\top \alpha^*)^2\right] \le \beta^\top (A^\top \Sigma_W^{-1} A)^{-1}\beta. \tag{9}$$

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 (7), for any $\widehat{B}$, that

$$\mathcal{R}(\widehat{B}) - \sigma^2 \geq \mathbb{E}_\theta \left[(Z_*^\top \beta - X_*^\top \alpha^*)^2 \right] \gtrsim \|\Sigma_W\|_{\mathrm{op}} \beta^\top (A^\top A)^{-1} \beta. \tag{10}$$

Therefore a risk analysis of linear predictors under factor regression models, which consists in upper bounding $\mathcal{R}(\widehat{B}) - \sigma^2$, 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 2 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 $\mathcal{R}(\widehat{B}) - \sigma^2$ to vanish as $p \to \infty$.

2.3 Upper Bound of the Risk $\mathcal{R}(\widehat{B})$

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 (3).

The prediction risk bound depends on W in Definition 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 22 in Appendix C. For any θ such that $(X, Y) \sim \text{sG-FRM}(\theta)$, one has

$$\mathbb{P}_\theta \left\{ \frac{1}{n} \|\boldsymbol{W}^\top \boldsymbol{W}\|_{\mathrm{op}} \leq \delta_W \right\} \geq 1 - e^{-n} \tag{11}$$

where δ_W is defined as

$$\delta_W := \delta_W(\theta) = c \left[\|\Sigma_W(\theta)\|_{\mathrm{op}} + \frac{\mathrm{tr}(\Sigma_W(\theta))}{n} \right], \tag{12}$$

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 $\lambda_1(\Sigma_W) \leq C$ for some constant $C > 0$, one has $\delta_W \lesssim 1 + p/n$. When $\lambda_p(\Sigma_W) \geq c$ for some constant $c > 0$, we have $\delta_W \gtrsim 1 + p/n$. In particular, if $c \leq \lambda_p(\Sigma_W) \leq \lambda_1(\Sigma_W) \leq C$, 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 $P_{\widehat{B}}^{\perp} = \boldsymbol{I}_p - P_{\widehat{B}}$ 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 15 in Appendix B, 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 $\widehat{B}$, $\boldsymbol{X}P_{\widehat{B}}^{\perp}$, is not used for prediction. It is natural therefore that the size of this component, as measured by its largest singular value, $\sigma_1^2(\boldsymbol{X}P_{\widehat{B}}^{\perp})$, 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 $\sigma_{\widehat{r}}^2(\boldsymbol{X}P_{\widehat{B}})$. These two quantities appear in the risk bound below.

We now state our main theorem; its proof is deferred to Appendix B.1.2. Recall that $\mathcal{R}(\widehat{B})$ is the risk defined in (4). Write $a \wedge b = \min\{a, b\}$.

Theorem 3 *Let $\widehat{B} = \widehat{B}(\boldsymbol{X}) \in \mathbb{R}^{p\times q}$ for some $q \geq 1$, and set*

$$\widehat{r} := \mathit{rank}\left(\boldsymbol{X}P_{\widehat{B}}\right), \qquad \widehat{\eta} := \frac{1}{n}\sigma_{\widehat{r}}^2\left(\boldsymbol{X}P_{\widehat{B}}\right), \qquad \widehat{\psi} := \frac{1}{n}\sigma_1^2\left(\boldsymbol{X}P_{\widehat{B}}^{\perp}\right). \tag{13}$$

For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \leq Cn/\log n$ 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

$$\begin{aligned}\mathbb{P}_\theta\Bigg\{\mathcal{R}(\widehat{B}) - \sigma^2 \;\lesssim\; &\left[\frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\eta}}\widehat{r} + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)(K \wedge \widehat{r} + \log n)\right]\frac{\sigma^2}{n} \\ &+ \left[\left(1 + \frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\eta}}\right)\delta_W + \left(1 + \frac{\delta_W}{\widehat{\eta}}\right)\widehat{\psi}\right]\beta^\top(A^\top A)^{-1}\beta\Bigg\} \geq 1 - c/n.\end{aligned} \tag{14}$$

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 (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 3 is thus to reduce the task of bounding $\mathcal{R}(\widehat{B})$ 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 3 holds for any estimator $\widehat{B} \in \mathbb{R}^{p\times q}$ that is constructed from $\boldsymbol{X}$ with any $q \geq 1$. We now explain the various terms in the bound (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

$$\begin{aligned}\widehat{Y}^*_{\widehat{B}} - Z_*^\top\beta &= X_*^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon} + X_*^\top\alpha^* - Z_*^\top\beta + X_*^\top\left[\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta - \alpha^*\right] \\ &= X_*^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon} + \left(X_*^\top\alpha^* - Z_*^\top\beta\right) + X_*^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{Z}\beta - \boldsymbol{X}\alpha^*) \\ &\quad + X_*^\top\left[\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X} - \boldsymbol{I}_p\right]\alpha^*.\end{aligned} \tag{15}$$

We discuss the four terms above one by one.

- The first term leads to the following variance term in (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 $\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\|_{\text{op}}^2 = \|P_{\widehat{B}}(\boldsymbol{X}P_{\widehat{B}})^+\|_{\text{op}}^2 \leq (n\widehat{\eta})^{-1}$. The two factors $\|\Sigma_W\|_{\text{op}}/\widehat{\eta}$ 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 (15) lead to the following term in (14), which can be interpreted as arising from the fact that Z_* and $\boldsymbol{Z}$ are not observed:

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

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

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

  is irreducible, as argued in (10), the term $\|\Sigma_W\|_{\text{op}}/\widehat{\eta}$ has been explained in the first term, and the inflation factor $\delta_W/\|\Sigma_W\|_{\text{op}}$ is due to the inflated noise level of $n^{-1}\|\boldsymbol{W}^\top\boldsymbol{W}\|_{\text{op}}$ compared to $\|\Sigma_W\|_{\text{op}}$.

- The fourth term in (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 $\Sigma := \text{Cov}(X)$. Also noting that $\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}$ is a projection matrix, the fourth term in (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 (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 4 *In light of the above discussion, we make two important remarks. First, to maintain a fast rate of the risk bound in (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 (14) simplifies to*

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

As $\widehat{r} = \text{rank}(\boldsymbol{X}P_{\widehat{B}})$ 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.

3. Analysis of Principal Component Regression Under the Factor Regression Model

In this section we use the general result, Theorem 3, to derive risk bounds for the popular Principal Component Regression (PCR) method. For any integer $1 \le k \le \mathrm{rank}(\boldsymbol{X})$, the PCR-predictor PCR-$k$ corresponds to taking $\widehat{B} = \boldsymbol{U}_k$, 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 $k > \mathrm{rank}(\boldsymbol{X})$. All the proofs of this section can be found in Appendix B.2.

Corollary 5 *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),*

$$\mathbb{P}_\theta\left\{\mathcal{R}(\boldsymbol{U}_k) - \sigma^2 \lesssim \widehat{B}(k)\right\} \ge 1 - cn^{-1} \tag{16}$$

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

$$\widehat{B}_1(k) := \left[\frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\lambda}_k}k + \left(1 + \frac{\delta_W}{\widehat{\lambda}_k}\right)(K \wedge k + \log n)\right]\frac{\sigma^2}{n} \tag{17}$$

$$\widehat{B}_2(k) := \left(\frac{\|\Sigma_W\|_{\mathrm{op}}}{\widehat{\lambda}_k}\delta_W + \delta_W + \widehat{\lambda}_{k+1}\right)\beta^\top(A^\top A)^{-1}\beta. \tag{18}$$

Corollary 5 follows immediately from the identities $\sigma_k^2(\boldsymbol{X}P_{\boldsymbol{U}_k}) = \sigma_k^2(\boldsymbol{X})$ and $\sigma_1^2(\boldsymbol{X}P_{\boldsymbol{U}_k}^\perp) = \sigma_{k+1}^2(\boldsymbol{X})$, and an application of Theorem 3 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 (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 7 in Appendix B.2). Nevertheless, (16) illustrates the effect of k and hints at the choice $k = \widehat{s}$ with

$$\widehat{s} = \max\left\{k \ge 0 :\ \widehat{\lambda}_k \ge C_0\delta_W\right\}. \tag{19}$$

Here δ_W is defined in (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 Bunea and Xiao (2015), but its role in PCR has received little attention. By definition, $\widehat{\lambda}_{\widehat{s}+1} < C_0\delta_W \le \widehat{\lambda}_{\widehat{s}}$, 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 (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 6 *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 (19) for any $C_0 > 1$,*

$$\mathbb{P}_\theta \left\{ \mathcal{R}(\boldsymbol{U}_{\widehat{s}}) - \sigma^2 \lesssim (K + \log n) \frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1} \beta \right\} \geq 1 - O(n^{-1}). \tag{20}$$

Remark 7

1. *We refer to the method analyzed in Corollary 6 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 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 6 thus covers the risk of PCR-K, that is, the risk of the PCR predictor corresponding to the true K of this θ.*

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 (19) relies on a theoretical order δ_W in (12), which depends on the unknown quantities $\|\Sigma_W\|_{\text{op}}$ and $\text{tr}(\Sigma_W)$. 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 Bing and Wegkamp (2019), 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 Bing and Wegkamp (2019), we define

$$\widetilde{s} := \underset{0\leq k\leq \bar{K}}{\arg\min}\, \widehat{v}_k^2, \quad \text{with} \quad \widehat{v}_k^2 := \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(k)}\|_F^2}{np - \mu_n k}, \quad \text{and} \quad \bar{K} := \left\lfloor \frac{\kappa}{1+\kappa} \frac{np}{\mu_n} \right\rfloor \wedge n \wedge p, \tag{21}$$

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 Bing and Wegkamp (2019, 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 (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 7 above.

The theoretical guarantees proved in Bing and Wegkamp (2019) are based on the assumption that $\boldsymbol{W}$ has i.i.d. entries with zero mean and bounded fourth moments. Proposition 8 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 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 8 *Let $\widetilde{s}$ be defined in (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), $\log p \leq cn$, $K \leq \bar{K}$ and*

$$r_e(\Sigma_W) \geq c'(n \wedge p) \tag{22}$$

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

$$\mathbb{P}_\theta\left\{\widetilde{s} \leq K, \quad \widehat{\lambda}_{\widetilde{s}} \gtrsim \delta_W, \quad \widehat{\lambda}_{\widetilde{s}+1} \lesssim \delta_W\right\} \geq 1 - O(1/n). \tag{23}$$

Condition $K \leq \bar{K}$ holds, for instance, if $K \leq c''(n \wedge p)$ with $c'' \leq \kappa/(2c_0(1+\kappa))$. We explain the connection between restriction (22) and the proposed choice of μ_n. Using elementary algebra, Bing and Wegkamp (2019, Theorem 6 and Proposition 7) proves the deterministic result

$$\left\{\frac{2\sigma_1^2(\boldsymbol{W})}{\|\boldsymbol{W}\|_F^2/(np)} \leq \mu_n\right\} \subseteq \{\widetilde{s} \leq K\}, \tag{24}$$

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 8 in Appendix B.2), the bound

$$\frac{2\sigma_1^2(\boldsymbol{W})}{\|\boldsymbol{W}\|_F^2/(np)} \lesssim np\frac{\delta_W}{\mathrm{tr}(\Sigma_W)} = p + \frac{np}{r_e(\Sigma_W)} \tag{25}$$

holds with probability larger than $1 - O(1/n)$. Thus, in view of (24) and (25), the event $\{\widetilde{s} \leq K\}$ holds with high probability as soon as $\mu_n > p + np/r_e(\Sigma_W)$. Under (22), we arrive at the choice $\mu_n = c_0(n+p)$ and, in turn, $\bar{K} = O(n \wedge p)$.

We note that (22) holds, for instance, in the commonly considered setting

$$0 < c' \leq \lambda_p(\Sigma_W) \leq \lambda_1(\Sigma_W) \leq C' < \infty, \tag{26}$$

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 (21), while our interest is in factor regression models with dimensions allowed to grow with n.

Proposition 8 in conjunction with Corollary 5 immediately leads to the following risk bound of PCR-$\widetilde{s}$. It coincides with the bound for PCR-$\widehat{s}$ in display (20) of Corollary 6.

Corollary 9 *Let $\widetilde{s}$ be defined in (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 $K \leq Cn/\log n$ such that (X, Y) follows sG-FRM(θ), $\log p \leq cn$, $K \leq \bar{K}$ and (22) holds, for some positive constants $c = c(\gamma_w)$ and $c' = c'(\gamma_w)$, we have*

$$\mathbb{P}_\theta \left\{ \mathcal{R}(\boldsymbol{U}_{\widetilde{s}}) - \sigma^2 \lesssim (K + \log n)\frac{\sigma^2}{n} + \delta_W \beta^\top (A^\top A)^{-1} \beta \right\} \geq 1 - O(n^{-1}). \tag{27}$$

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 (10) above, which shows that a necessary condition for prediction with vanishing excess risk, under factor regression models with well conditioned Σ_W, is $\|\Sigma_W\|_{\text{op}} \beta^\top (A^\top A)^{-1} \beta \to 0$, which can be met when $p \to \infty$, as explained below.

This phenomenon was first quantified in Stock and Watson (2002), where it is shown that

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

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 Bai (2003) and Fan et al. (2013), 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 E.

In addition to being asymptotic in nature, the results in Stock and Watson (2002), and also those regarding the in-sample prediction accuracy, are established under the following

set of conditions: $K = O(1)$, $\|\beta\|^2 = O(1)$, $\|\Sigma_W\|_{\text{op}} = O(1)$, as $p \to \infty$, and

$$\frac{1}{p}A^\top A \to \boldsymbol{I}_K, \text{ as } p \to \infty, \quad \Sigma_Z \text{ is a diagonal matrix with distinct diagonal entries.} \quad (28)$$

These conditions serve as identifiability conditions for $\theta = (K, \beta, A, \Sigma_Z, \Sigma_W, \sigma^2)$ (Stock and Watson, 2002). Condition (28) further implies that, for some constants $0 < c \le C < \infty$,

$$p \lesssim \lambda_K(AA^\top) \le \lambda_1(AA^\top) \lesssim p, \quad c \le \lambda_K(\Sigma_Z) \le \lambda_1(\Sigma_Z) \le C. \quad (29)$$

In contrast, our Corollaries 5, 6 and 9 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 (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 $\lambda_K(\Sigma_Z) \ge c$. Then from Remark 7 we only require a condition much weaker than $\lambda_K(AA^\top) \gtrsim p$ of (Stock and Watson, 2002) given in (29) above, namely

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

Finally, the results in Stock and Watson (2002) 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 (28).

We conclude our comparison by giving the bound implied by our Corollary 6, should the more stringent conditions (29) be met. Since (29) implies that $\widehat{s} = K$ with high probability from Remark 7, Corollary 6 immediately yields, with probability $1 - O(n^{-1})$,

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

and thus, as in Stock and Watson (2002),

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

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

4. Analysis of Alternative Prediction Methods

In this section we illustrate the usage of the main Theorem 3 to derive risk bounds under a factor regression model for two other prediction methods: Generalized Least Squares (Bunea et al., 2020), 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 Bing et al. (2019). All proofs for this section are contained in Appendix B.3.

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

In the recent paper Bunea et al. (2020), risk bounds were established under the factor regression model for the Generalized Least Squares (GLS) predictor, which corresponds to taking $\widehat{B} = \boldsymbol{I}_p$:

$$\widehat{Y}^*_{\boldsymbol{I}_p} = X_*^\top \boldsymbol{X}^+ \boldsymbol{Y}. \tag{30}$$

We recover as these results in Corollary 10 and Corollary 11 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 3 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 Vershynin, 2012, Theorem 5.39) to show $\sigma_p^2(\boldsymbol{X}) \gtrsim \lambda_p(\Sigma_W)n$, which implies $\widehat{\eta} \gtrsim \lambda_p(\Sigma_W)$. Recall that $\kappa(\Sigma_W) \coloneqq \lambda_1(\Sigma_W)/\lambda_p(\Sigma_W)$.

Corollary 10 (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*

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

When p is much larger than n, the GLS becomes the minimum ℓ_2 norm interpolator (Bunea et al., 2020), one method studied in the recent wave of literature on the generalization of overparameterized models with zero or near-zero training error (Montanari et al., 2019; Bunea et al., 2020; Muthukumar et al., 2019, 2020; Hastie et al., 2019; Feldman, 2019; Belkin et al., 2019a,b, 2018a,b,c; Bartlett et al., 2019; Liang and Rakhlin, 2018). Theorem 3 can also be applied to recover a slightly modified form of the prediction risk bound from Bunea et al. (2020) 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 11 (GLS: high-dimensional setting. Interpolating predictors.) *For any $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ with $K \le Cn/\log n$ such that $(X, Y) \sim$ sG-FRM(θ), suppose $\widetilde{W}$ defined in Definition 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*

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

By Proposition 6 of Bunea et al. (2020), we have $\sigma_n^2(\boldsymbol{X}) \gtrsim \mathrm{tr}(\Sigma_W)$ with high probability when $r_e(\Sigma_W) \gtrsim n$. Corollary 11 thus follows from Theorem 3 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 11, together with a comparison with PCR-k prediction, is presented in Section 4.3.

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 Bing et al. (2019) 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 (1), one further assumes the following model specifications.

Assumption 1

(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 (Bing et al., 2020). In fact, $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ can be further shown to be identifiable (Bing et al., 2019).

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 (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 (1) implies

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

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

$$\widetilde{Z} = \mathrm{Cov}(Z, \bar{X})[\mathrm{Cov}(\bar{X})]^{-1} \bar{X} = \Sigma_Z \left(\Sigma_Z + A^+ \Sigma_W A^{+\top} \right)^{-1} A^+ X. \tag{31}$$

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 (3) with $\widehat{B} = \widehat{A}$, with $\widehat{A}$ being an estimator of A tailored to the ER model, developed in Bing et al. (2020). We summarize the construction of $\widehat{A}$ in Appendix D 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 2 *Assume $c \leq \lambda_K(\Sigma_Z) \leq \lambda_1(\Sigma_Z) \leq C$ 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 3, 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 Bing et al. (2020),

$$\|\widehat{A} - A\|^2_{\text{op}} \leq \|A_J\|_0 \log(n \vee p)/n \tag{32}$$

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 (32) can be leveraged to obtain a small improvement in the risk bound by slightly adjusting the proof of Theorem 3. 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 12 (Prediction in Essential Regression) *Suppose $(X, Y) \sim$ sG-FRM(θ) with $\theta = (K, A, \beta, \Sigma_Z, \Sigma_W, \sigma^2)$ satisfying Assumptions 1 & 2, $K \leq Cn/\log n$ and*

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

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

$$\mathcal{R}(\widehat{A}) - \sigma^2 \lesssim \frac{K + \log n}{n}\sigma^2 + \psi_n(A_J)\beta^\top (A^\top A)^{-1}\beta. \tag{33}$$

Remark 13

1. *We note that the bound (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 1, has been used in Bing et al. (2020) 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 $\mathcal{R}(\widehat{A})$ as the risk of predicting Y from predicted cluster centers, on the basis of data that exhibits a latent cluster structure with overlap.*

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 6, Remark 7, Corollary 11 and Theorem 12, 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

$$\begin{aligned}
\mathcal{R}(\boldsymbol{U}_K) - \sigma^2 &\lessapprox \frac{K}{n}\sigma^2 + \frac{K}{p}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 + \frac{K}{n}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 \\
\mathcal{R}(\widehat{A}) - \sigma^2 &\lessapprox \frac{K}{n}\sigma^2 + \frac{K}{p}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 + \frac{K\|A_J\|_0}{np}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2 \qquad (34) \\
\mathcal{R}(\boldsymbol{I}_p) - \sigma^2 &\lessapprox \frac{K}{n}\sigma^2 + \frac{n}{p}\sigma^2 + \frac{K}{n}\|\Sigma_W\|_{\mathrm{op}} R_\beta^2
\end{aligned}$$

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 $\mathcal{R}(\boldsymbol{U}_K)$ and $\mathcal{R}(\widehat{A})$, and note that the difference in their respective errors bounds depends on the sparsity of A_J. The risk bound on $\mathcal{R}(\boldsymbol{U}_K)$ is valid for any θ such that $(X, Y) \sim \text{sG-FRM}(\theta)$, 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 $\mathcal{R}(\widehat{A})$ has a faster rate. This advantage becomes considerable if $|J| = o(p)$, that is, in the presence 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 $\mathcal{R}(\widehat{A})$ has a slower rate than $\mathcal{R}(\boldsymbol{U}_K)$. 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

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 1 into account.

is the main premise of Essential Regression. Furthermore, in this risk bound comparison, $\mathcal{R}(\widehat{A})$ 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 3 above. The risk bound above will then hold under conditions discussed in Remark 7.

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 $\mathcal{R}(\boldsymbol{I}_p)$ 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.

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

$$\widehat{m} \coloneqq \arg\min_{m\in[M]} \sum_{i\in D_2} (Y_i - X_i^\top \widehat{\alpha}_m)^2, \tag{35}$$

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 $\widehat{\alpha} \coloneqq \widehat{\alpha}_{\widehat{m}}$ as our predictor, for which we establish the following oracle inequality, which is an adaptation of Theorem 2.1 from Wegkamp (2003) 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 Wegkamp (2003). The proof is deferred to Appendix B.4.

Theorem 14 *Let $\widehat{\alpha} \coloneqq \widehat{\alpha}_{\widehat{m}}$, where $\widehat{m}$ is defined in (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 $a > 0$,*

$$\mathbb{P}_\theta\Bigg\{\mathcal{R}(\widehat{\alpha}) - \sigma^2 \le (1+a)^2 \min_{m\in[M]}\{\mathcal{R}(\widehat{\alpha}_m) - \sigma^2\} + C(a)\left(\sigma^2 \vee \max_{m\in[M]}\{\mathcal{R}(\widehat{\alpha}_m) - \sigma^2\}\right)\frac{\log(nM)}{n}\Bigg\} \ge 1 - c'n^{-1}, \tag{36}$$

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

In the bound above, the worst excess risk $\max_m\{\mathcal{R}(\widehat{\alpha}_m) - \sigma^2\}$ 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 (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 4.3. If the three excess risks in (34) are all $O(1)$, which is met under the conditions discussed in detail in Section 4.3, then the bound (36) becomes

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

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

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:

- PCR-$\widetilde{s}$ with $\widetilde{s}$ obtained from (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 Lam and Yao (2012); Ahn and Horenstein (2013); [4]
- GLS: the Generalized Least Squares predictor defined in (30);
- ER-A: the Essential Regression predictor with $\widehat{B} = \widehat{A}$ in (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;

4. We have also implemented the selection criterion suggested by Bai and Ng (2002), but it had inferior performance, and is for this reason not included in our comparison here.

- MS: the selected predictor from (35) in Section 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 (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 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 $\mathrm{Unif}(1, 3)$, and β is generated with entries sampled from $\mathrm{Unif}(0, 3)$.

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 1. 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 $\mathrm{Unif}(0, 1/s_j)$ 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 $\|\widehat{\boldsymbol{Y}}_{new} - \boldsymbol{Z}_{new}\beta\|^2/n$.

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 (8), one at a time. The MS predictor is based on (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 1. Since both PCR-$\widetilde{s}$ and PCR-ratio consistently select the true K, we only present the result for PCR-K.

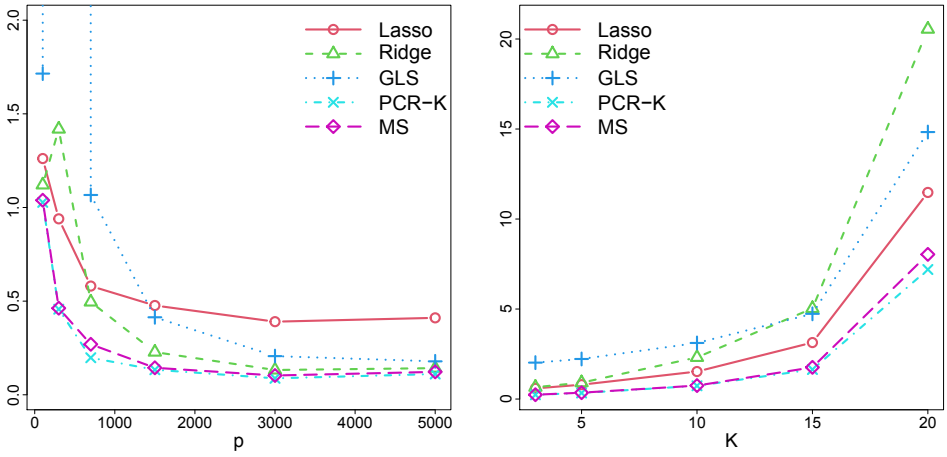

Figure 1: Prediction risks of different predictors under the factor regression model as p and K vary separately

Results: Overall, it is clear that the MS predictor selects the best predictor in almost all settings, corroborating Theorem 14. 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 exceptions 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 2 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, Stock and Watson (2002); Bai (2003); Bai and Ng (2008, 2006); Fan et al. (2013) 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 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 (Hastie et al., 2019). 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 Hastie et al. (2019), 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 3 and 4. We also note that the performance of Ridge deteriorates faster than the other methods when K grows.

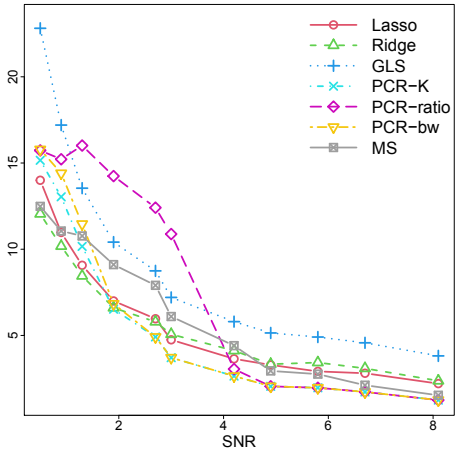

Figure 2: Prediction risks of different predictors under the factor regression model as SNR varies

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 α, we calculate the SNR and plot the prediction risks of each predictor in Figure 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 under-estimate 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 9 and Remark 7.

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 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 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.

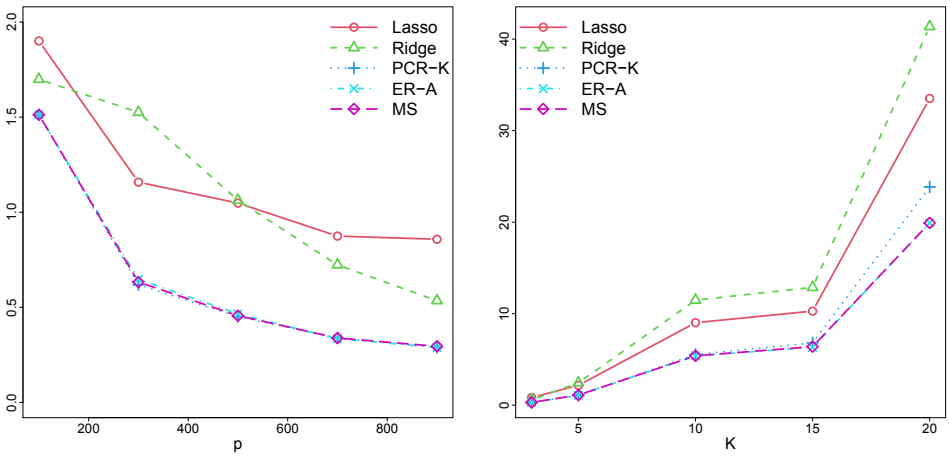

Figure 3: Prediction risks of different predictors under the Essential Regression model as p and K vary separately

Acknowledgments

Acknowledgements Bunea and Wegkamp are supported in part by NSF grants DMS-1712709 and DMS-2015195.

Appendix A. Organization of Appendices

We provide section-by-section proofs for the main results in Appendices B.1—B.4. Auxiliary lemmas are collected in Appendix C. Appendix D contains the procedure of estimating A under the Essential Regression framework while comparison with more existing literature on factor models is stated in Appendix E.

Appendix B. 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 15 *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 $r_0 = \mathrm{rank}(\widehat{B})$. 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.1 Proofs for Section 2

B.1.1 Proof of Lemma 2

Let $\Sigma_X = \mathrm{Cov}(X)$, $\Sigma_{XY} = \mathrm{Cov}(X, Y)$. 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 $\alpha^* = \Sigma_X^{-1}\Sigma_{XY}$,

$$\mathcal{R}^* - \sigma^2 = \mathbb{E}[(X^\top\alpha^* - Z^\top\beta)^2]. \tag{37}$$

Using this expression, and the factor model structure $X = AZ + W$, $Y = Z^\top\beta + \varepsilon$, the proof of Lemma 4 in Bunea et al. (2020) uses the Woodbury matrix identity to simplify (37), arriving at

$$\mathcal{R}^* - \sigma^2 = \beta^\top(\Sigma_Z^{-1} + A^\top\Sigma_W^{-1}A)^{-1}\beta.$$

Letting $H = \Sigma_Z^{1/2}A^\top\Sigma_W^{-1}A\Sigma_Z^{1/2}$, we then have

$$\begin{aligned}\mathcal{R}^* - \sigma^2 &= \beta^\top\Sigma_Z^{1/2}(\boldsymbol{I}_K + H)^{-1}\Sigma_Z^{1/2}\beta \\ &= \beta^\top\Sigma_Z^{1/2}H^{-1/2}(\boldsymbol{I}_K + H^{-1})^{-1}H^{-1/2}\Sigma_Z^{1/2}\beta.\end{aligned}$$

To obtain the upper bound on $\mathcal{R}^*$ we use

$$\mathcal{R}^*-\sigma^2 = \beta^\top\Sigma_Z^{1/2}H^{-1/2}(\boldsymbol{I}_K+H^{-1})^{-1}H^{-1/2}\Sigma_Z^{1/2}\beta \leq \frac{\beta^\top\Sigma_Z^{1/2}H^{-1}\Sigma_Z^{1/2}\beta}{1+\lambda_K(H^{-1})} \leq \beta^\top(A^\top\Sigma_W^{-1}A)^{-1}\beta,$$

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

$$\mathcal{R}^*-\sigma^2 = \beta^\top\Sigma_Z^{1/2}H^{-1/2}(\boldsymbol{I}_K+H^{-1})^{-1}H^{-1/2}\Sigma_Z^{1/2}\beta \geq \frac{\beta^\top\Sigma_Z^{1/2}H^{-1}\Sigma_Z^{1/2}\beta}{1+\|H^{-1}\|_{\mathrm{op}}} = \frac{\beta^\top(A^\top\Sigma_X^{-1}A)^{-1}\beta}{1+\lambda_K^{-1}(H)}.$$

Furthermore,

$$\lambda_K(H) = \lambda_K(\Sigma_Z^{1/2}A^\top\Sigma_W^{-1}A\Sigma_Z^{1/2}) \geq \lambda_K(A\Sigma_Z A^\top)/\|\Sigma_W\|_{\mathrm{op}} = \xi,$$

so using this in the previous display,

$$\mathcal{R}^* - \sigma^2 \geq \frac{\beta^\top(A^\top\Sigma_X^{-1}A)^{-1}\beta}{1+\xi^{-1}} = \frac{\xi}{1+\xi}\cdot\beta^\top(A^\top\Sigma_X^{-1}A)^{-1}\beta,$$

as claimed. ■

B.1.2 Proof of Theorem 3

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 (3). Pick any θ with $K \leq (Cn/\log n)\wedge p$ such that (X, Y) follows $\mathrm{FRM}(\theta)$ 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

$$\begin{aligned}
\mathcal{R}(\widehat{B}) - \sigma^2 &= \mathbb{E}_{(Z_*, W_*)}\left[\left(\widehat{Y}^*_{\widehat{B}} - Z_*^\top \beta\right)^2\right] \\
&= \mathbb{E}_{Z_*}\left[\left(Z_*^\top A^\top \widehat{\alpha}_{\widehat{B}} - Z_*^\top \beta\right)^2\right] + \mathbb{E}_{W_*}\left[\left(W_*^\top \widehat{\alpha}_{\widehat{B}}\right)^2\right] \qquad (38) \\
&= \left\|\Sigma_Z^{1/2}\left(A^\top \widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2 + \left\|\Sigma_W^{1/2}\widehat{\alpha}_{\widehat{B}}\right\|^2 \\
&\le \left\|\Sigma_Z^{1/2}\left(A^\top \widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2 + \|\Sigma_W\|_{\mathrm{op}}\left\|\widehat{\alpha}_{\widehat{B}}\right\|^2. \qquad (39)
\end{aligned}$$

We define an event $\mathcal{E}^*$ in (40) below, on which we bound the risk. Invoking Lemmas 17, 18 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 16, $\mathbb{P}(\mathcal{E}^*) \ge 1 - cn^{-1}$, which completes the proof. ■

We state and prove three lemmas which are used in the proof of Theorem 3. Recall that

$$\widehat{r} = \mathrm{rank}(\boldsymbol{X}P_{\widehat{B}}), \qquad \widehat{\psi} = \frac{1}{n}\sigma_1^2\left(\boldsymbol{X}P^\perp_{\widehat{B}}\right), \qquad \widehat{\eta} = \frac{1}{n}\sigma_{\widehat{r}}^2\left(\boldsymbol{X}P_{\widehat{B}}\right).$$

Lemma 16 *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*

$$\mathcal{E}^* := \mathcal{E}_{\boldsymbol{Z}} \cap \mathcal{E}_{\boldsymbol{W}} \cap \mathcal{E}'_{\boldsymbol{W}} \cap \mathcal{E}_M \cap \mathcal{E}_{M'} \cap \mathcal{E}_{\boldsymbol{Z}\beta}. \qquad (40)$$

Here, for some constants $c(\gamma_z)$ and $c'(\gamma_w)$ depending on γ_z and γ_w, respectively,

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{Z}} &:= \left\{\lambda_K\left(\Omega^{1/2}\frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{Z}\Omega^{1/2}\right) \ge c(\gamma_z)\right\}, \\
\mathcal{E}_{\boldsymbol{Z}\beta} &:= \left\{\frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 \le 8\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta + 2\widehat{\psi}\beta^\top(A^\top A)^{-1}\beta\right\}, \\
\mathcal{E}_{\boldsymbol{W}} &:= \left\{\frac{1}{n}\left\|\boldsymbol{W}^\top\boldsymbol{W}\right\|_{\mathrm{op}} \le \delta_W\right\}, \\
\mathcal{E}'_{\boldsymbol{W}} &:= \left\{\frac{1}{n}\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 \le 4\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta\right\}, \\
\mathcal{E}_M &:= \left\{\boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon} \le 2\gamma_\varepsilon^2\sigma^2\left[2\|M\|_{\mathrm{op}}\log n + \mathrm{tr}(M)\right]\right\}, \\
\mathcal{E}_{M'} &:= \left\{\boldsymbol{\varepsilon}^\top M'\boldsymbol{\varepsilon} \le 2\gamma_\varepsilon^2\sigma^2\left[2\|M'\|_{\mathrm{op}}\log n + \mathrm{tr}(M')\right]\right\},
\end{aligned}$$

with $\Omega := \Sigma_Z^{-1}$, δ_W defined in (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 Vershynin (2012) and $K \log n \leq C(\gamma_z)n$, we find $\mathbb{P}\{\mathcal{E}_{\boldsymbol{Z}}^c\} \lesssim n^{-c'K}$. From Lemma 22 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 21 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

$$\mathbb{P}\{(\mathcal{E}'_{\boldsymbol{W}})^c\} = \mathbb{P}\left\{\frac{1}{n}\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 > 4\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta\right\} \leq n^{-1}. \tag{41}$$

We prove $\mathcal{E}'_{\boldsymbol{W}} \cap \mathcal{E}_{\boldsymbol{Z}\beta} = \mathcal{E}_{\boldsymbol{W}'}$ in Lemma 19. 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 21 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

$$\mathbb{P}\{\mathcal{E}^{*c}\} \leq \mathbb{P}\left\{\mathcal{E}_{\boldsymbol{Z}}^c\right\} + \mathbb{P}\left\{\mathcal{E}_{\boldsymbol{W}}^c\right\} + \mathbb{P}\left\{(\mathcal{E}'_{\boldsymbol{W}})^c\right\} + \mathbb{P}\{\mathcal{E}_M^c\} + \mathbb{P}\{\mathcal{E}_{M'}^c\} \lesssim n^{-1}.$$

■

Lemma 17 *Under conditions of Theorem 3, on the event $\mathcal{E}^*$ defined in (40),*

$$\|\widehat{\alpha}_{\widehat{B}}\|^2 \lesssim_\theta \frac{(\widehat{r} + \log n)\sigma^2}{n\widehat{\eta}} + \beta^\top(A^\top A)^{-1}\beta + \widehat{\eta}^{-1}\left(\widehat{\psi}\beta^\top(A^\top A)^{-1}\beta + \beta^\top A^+\Sigma_W A^{+\top}\beta\right). \tag{42}$$

Proof Starting with the identity

$$\widehat{\alpha}_{\widehat{B}} = \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Y} = \widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{Z}\beta + \boldsymbol{\varepsilon}), \tag{43}$$

with $(\boldsymbol{X}\widehat{B})^+ \coloneqq (\widehat{B}\boldsymbol{X}^\top\boldsymbol{X}\widehat{B})^+\widehat{B}^\top\boldsymbol{X}^\top$, we have

$$\|\widehat{\alpha}_{\widehat{B}}\|^2 \leq 2\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 + 2\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2.$$

To bound the first term, notice that

$$\begin{aligned}\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 &= \boldsymbol{\varepsilon}^\top(\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon} \\ &= \boldsymbol{\varepsilon}^\top M\boldsymbol{\varepsilon} \\ &\leq 2\gamma_\varepsilon^2\sigma^2\Big[2\|M\|_{\mathrm{op}}\log n + \mathrm{tr}(M)\Big],\end{aligned}$$

where the last step holds on $\mathcal{E}^*$ (in particular, on $\mathcal{E}_M \subset \mathcal{E}^*$). Observe that, on $\mathcal{E}^*$,

$$\begin{aligned}\mathrm{tr}(M) &= \mathrm{tr}\left((\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\right) \\ &\leq \mathrm{rank}(\boldsymbol{X}\widehat{B}) \cdot \|M\|_{\mathrm{op}} \\ &= \widehat{r}\|M\|_{\mathrm{op}}.\end{aligned}$$

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 $\mathcal{E}^*$,

$$\begin{aligned}
\|M\|_{\mathrm{op}} &= \left\|(\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}} \\
&\overset{(i)}{=} \left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{X}\widehat{B})^{+\top}\widehat{B}^\top\right\|_{\mathrm{op}} \\
&= \left\|\widehat{B}(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B})^+\widehat{B}\boldsymbol{X}^\top\boldsymbol{X}\widehat{B}(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B})^+\widehat{B}^\top\right\|_{\mathrm{op}} \\
&= \left\|\widehat{B}\left(\widehat{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\widehat{B}\right)^+\widehat{B}^\top\right\|_{\mathrm{op}} \\
&= \left\|U\left(U^\top\boldsymbol{X}^\top\boldsymbol{X}U\right)^+U^\top\right\|_{\mathrm{op}} \\
&\overset{(ii)}{\leq} \sigma_{\widehat{r}}^{-2}(\boldsymbol{X}U) \\
&\overset{(iii)}{=} (n\widehat{\eta})^{-1}
\end{aligned} \tag{44}$$

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 $\mathcal{E}^*$,

$$\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 \leq \frac{2\gamma_\varepsilon^2\sigma^2}{n\widehat{\eta}}(\widehat{r} + 2\log n). \tag{45}$$

On the other hand, by $A^\top A^{+\top} = \boldsymbol{I}_K$ and $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$, observe that

$$\begin{aligned}
\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} &= \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}A^\top A^{+\top} \\
&= \widehat{B}(\boldsymbol{X}\widehat{B})^+(\boldsymbol{X} - \boldsymbol{W})A^{+\top} \\
&= \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}A^{+\top} + \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}^\perp A^{+\top} - \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{W}A^{+\top}.
\end{aligned} \tag{46}$$

By $P_{\widehat{B}} = \widehat{B}\widehat{B}^+$ and the inequality $(a+b+c)^2 \leq 3a^2 + 3b^2 + 3c^2$,

$$\begin{aligned}
\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2 &\leq 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+A^{+\top}\beta\right\|^2 + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}P_{\widehat{B}}^\perp A^{+\top}\beta\right\|^2 \\
&\quad + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{W}A^{+\top}\beta\right\|^2 \\
&\leq 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{X}\widehat{B}\widehat{B}^+\right\|_{\mathrm{op}}^2\left\|A^{+\top}\beta\right\|^2 + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\left\|\boldsymbol{X}P_{\widehat{B}}^\perp\right\|_{\mathrm{op}}^2\left\|A^{+\top}\beta\right\|^2 \\
&\quad + 3\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\mathrm{op}}^2\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2.
\end{aligned} \tag{47}$$

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 (44), concludes

$$\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2 \le 3\beta^\top(A^\top A)^{-1}\beta + 3\widehat{\eta}^{-1}\left(\widehat{\psi}\beta^\top(A^\top A)^{-1}\beta + 4\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta\right). \quad (48)$$

Collecting (45)—(48) concludes the proof. ■

Lemma 18 *Under conditions of Theorem 3, on the event $\mathcal{E}^*$ defined in (40),*

$$\begin{aligned}\left\|\Sigma_Z^{1/2}\left(A^\top\widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2 \lesssim_\theta &\left(1+\frac{\delta_W}{\widehat{\eta}}\right)\left(\frac{K\wedge\widehat{r}+\log n}{n}\sigma^2 + \beta^\top A^+\Sigma_W A^{+\top}\beta\right)\\ &+\left[\left(1+\frac{\delta_W}{\widehat{\eta}}\right)\widehat{\psi}+\delta_W\right]\beta^\top(A^\top A)^{-1}\beta.\end{aligned}$$

Proof Use identity (43) and the inequality $(x+y)^2 \le 2x^2+2y^2$ to find

$$\begin{aligned}&\left\|\Sigma_Z^{1/2}\left(A^\top\widehat{\alpha}_{\widehat{B}} - \beta\right)\right\|^2\\ &\le 2\left\|\Sigma_Z^{1/2}[A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2 + 2\left\|\Sigma_Z^{1/2}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2. \quad (49)\end{aligned}$$

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

$$\begin{aligned}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+ - \boldsymbol{Z}^+ &= \boldsymbol{Z}^+\boldsymbol{Z}A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+ - \boldsymbol{Z}^+ \qquad (\text{by } \boldsymbol{Z}^+\boldsymbol{Z} = \boldsymbol{I}_K \text{ on } \mathcal{E}^*)\\ &= \boldsymbol{Z}^+(\boldsymbol{X}-\boldsymbol{W})\widehat{B}(\boldsymbol{X}\widehat{B})^+ - \boldsymbol{Z}^+\\ &= -\boldsymbol{Z}^+P^\perp_{\boldsymbol{X}\widehat{B}} - \boldsymbol{Z}^+\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+, \quad (50)\end{aligned}$$

which yields

$$\begin{aligned}&\left\|\Sigma_Z^{1/2}[A^\top\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2\\ &\le 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 + 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2\\ &\lesssim \frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 + \frac{1}{n}\left\|\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2 \quad (51)\\ &\lesssim \frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 + \delta_W\cdot\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z}\beta\right\|^2.\end{aligned}$$

We used $\|\Sigma_Z^{1/2}\boldsymbol{Z}^+\|_{\rm op} = \sigma_K^{-1}(\boldsymbol{Z}\Omega^{-1/2}) \lesssim 1/\sqrt{n}$ on $\mathcal{E}^*$ in the third line. The event $\mathcal{E}_{\boldsymbol{Z}\beta}$ and (48) conclude

$$\begin{aligned}&\left\|\Sigma_Z^{1/2}[A^\top \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{Z} - \boldsymbol{I}_K]\beta\right\|^2 \\ &\quad\lesssim \left(1+\frac{\delta_W}{\widehat{\eta}}\right)\left(\beta^\top A^+\Sigma_W A^{+\top}\beta + \widehat{\psi}\beta^\top(A^\top A)^{-1}\beta\right) + \delta_W\beta^\top(A^\top A)^{-1}\beta.\end{aligned} \tag{52}$$

For the second term in (49), we use that on $\mathcal{E}^*$ (in particular, $\mathcal{E}_{M'} \subset \mathcal{E}^*$),

$$\left\|\Sigma_Z^{1/2}A^\top \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 \le 2\gamma_\varepsilon^2\sigma^2\left[2\|M'\|_{\rm op}\log n + \mathrm{tr}(M')\right]$$

Since $\mathrm{rank}(\Sigma_Z) = K$ and $\mathrm{rank}(\boldsymbol{X}\widehat{P}_{\widehat{B}}) = \widehat{r}$, we have

$$\mathrm{tr}(M') \le (K \wedge \widehat{r}\,)\|M'\|_{\rm op}.$$

Moreover,

$$\begin{aligned}\|M'\|_{\rm op} = \left\|\Sigma_Z^{1/2}A^\top \widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\rm op}^2 &\le 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+P_{\boldsymbol{X}\widehat{B}}\right\|_{\rm op}^2 + 2\left\|\Sigma_Z^{1/2}\boldsymbol{Z}^+\boldsymbol{W}\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\rm op}^2 \\ &\lesssim \frac{1}{n} + \delta_W\cdot\left\|\widehat{B}(\boldsymbol{X}\widehat{B})^+\right\|_{\rm op}^2\end{aligned}$$

by using (50) in the first line and $\mathcal{E}^*$ in the second line. Invoking (44) concludes that, on $\mathcal{E}^*$,

$$\left\|\Sigma_Z^{1/2}A^\top \widehat{B}(\boldsymbol{X}\widehat{B})^+\boldsymbol{\varepsilon}\right\|^2 \lesssim \frac{(K\wedge\widehat{r}+\log n)\sigma^2}{n}\left(1+\frac{\delta_W}{\widehat{\eta}}\right). \tag{53}$$

Plugging (52) and (53) into (49) completes the proof. ■

Lemma 19 *Under conditions of Theorem 3, on the event $\mathcal{E}'_{\boldsymbol{W}}$ from (40),*

$$\frac{1}{n}\left\|P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 \le 8\gamma_w^2\beta^\top A^+\Sigma_W A^{+\top}\beta + 2\widehat{\psi}\beta^\top(A^\top A)^{-1}\beta. \tag{54}$$

Proof By $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$, one has

$$\begin{aligned}P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta &= P^\perp_{\boldsymbol{X}\widehat{B}}\left(\boldsymbol{X}A^{+\top} - \boldsymbol{W}A^{+\top}\right)\beta \\ &= -P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta + P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}A^{+\top}\beta \\ &= -P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta + P^\perp_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}\left(A^{+\top} - \widehat{B}G\right)\beta\end{aligned}$$

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

$$P^{\perp}_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta = P^{\perp}_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta + P^{\perp}_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}P^{\perp}_{\widehat{B}}A^{+\top}\beta.$$

Then by the basic inequality $(a+b)^2 \le 2a^2 + 2b^2$,

$$\begin{aligned}\left\|P^{\perp}_{\boldsymbol{X}\widehat{B}}\boldsymbol{Z}\beta\right\|^2 &\le 2\left\|P^{\perp}_{\boldsymbol{X}\widehat{B}}\boldsymbol{W}A^{+\top}\beta\right\|^2 + 2\left\|P^{\perp}_{\boldsymbol{X}\widehat{B}}\boldsymbol{X}P^{\perp}_{\widehat{B}}A^{+\top}\beta\right\|^2 \\ &\le 2\left\|P^{\perp}_{\boldsymbol{X}\widehat{B}}\right\|^2_{\mathrm{op}}\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 + 2\left\|\boldsymbol{X}P^{\perp}_{\widehat{B}}\right\|^2_{\mathrm{op}}\left\|A^{+\top}\beta\right\|^2 \\ &\le 2\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 + 2n\widehat{\psi}\beta^{\top}(A^{\top}A)^{-1}\beta\end{aligned} \tag{55}$$

where we invoked the definition of $\widehat{\psi}$ in the last line. Invoke $\mathcal{E}'_{\boldsymbol{W}}$ from (40) to finish the proof. ■

B.2 Proofs for Section 3

B.2.1 Proof of Corollary 5

The corollary is an application of Theorem 3 with $\widehat{B} = \boldsymbol{U}_k$. 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

$$\begin{aligned}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\top} &= \sum_{1\le j\le k}\boldsymbol{D}_{jj}\boldsymbol{V}_{\bullet j}\boldsymbol{U}_{\bullet j}^{\top} + \sum_{j>k}\boldsymbol{D}_{jj}\boldsymbol{V}_{\bullet j}\boldsymbol{U}_{\bullet j}^{\top} \\ &\coloneqq \boldsymbol{V}_k\boldsymbol{D}_k\boldsymbol{U}_k^{\top} + \boldsymbol{V}_{(-k)}\boldsymbol{D}_{(-k)}\boldsymbol{U}_{(-k)}^{\top}.\end{aligned}$$

The diagonal matrix $\boldsymbol{D}$ contains the non-increasing singular values and $\boldsymbol{U}_k$ contains the corresponding k right-singular vectors. Consequently,

$$\begin{aligned}&\mathrm{rank}(\boldsymbol{X}\boldsymbol{U}_k) = \mathrm{rank}(\boldsymbol{V}_k\boldsymbol{D}_k) = k, \\ &\sigma_1^2\left(\boldsymbol{X}P^{\perp}_{\boldsymbol{U}_k}\right) = \left\|\boldsymbol{X}\boldsymbol{U}_{(-k\ )}\boldsymbol{U}_{(-k)}^{\top}\right\|^2_{\mathrm{op}} = \left\|\boldsymbol{V}_{(-k\ )}\boldsymbol{D}_{(-k)}\boldsymbol{U}_{(-k\ )}^{\top}\right\|^2_{\mathrm{op}} = \sigma^2_{k+1}(\boldsymbol{X}) = n\widehat{\lambda}_{k+1}, \\ &\sigma_1^2(\boldsymbol{X}P_{\boldsymbol{U}_k}) = \sigma_1^2\left(\boldsymbol{V}_k\boldsymbol{D}_k\boldsymbol{U}_k^{\top}\right) = \sigma_k^2(\boldsymbol{X}) = n\widehat{\lambda}_k.\end{aligned}$$

Invoke Theorem 3 with $\widehat{B} = \boldsymbol{U}_k$, $\widehat{r} = k$, $\widehat{\psi} = \widehat{\lambda}_{k+1}$ and $\widehat{\eta} = \widehat{\lambda}_k$ to conclude the proof.■

B.2.2 Proof of Corollary 6 & Remark 7

We first prove Corollary 6. From Corollary 5, it suffices to show $\mathbb{P}_\theta\{\widehat{s} \le K\} \ge 1 - c/n$, which is guaranteed by proving

$$\mathbb{P}_\theta\left\{\frac{1}{n}\sigma^2_{K+1}(\boldsymbol{X}) < C_0\delta_W\right\} \ge 1 - c/n.$$

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 (11) and $C_0 > 1$. ■

To prove Remark 7, we will show

$$\mathbb{P}\left\{\widehat{\lambda}_K \gtrsim \lambda_k(A\Sigma_Z A^\top) - \delta_W\right\} \geq 1 - n^{-c}.$$

Note that Weyl's inequality yields

$$\sigma_k(\boldsymbol{X}) \geq \sigma_k(\boldsymbol{Z}A^\top) - \sigma_1(\boldsymbol{W}) \geq \sigma_K(\boldsymbol{Z}\Sigma_Z^{-1/2})\sigma_k(\Sigma_Z^{1/2}A^\top) - \sigma_1(\boldsymbol{W}).$$

We obtain the desired result by invoking $\mathcal{E}_{\boldsymbol{Z}}$ from Lemma 16 and (11). ■

B.2.3 Proof of Proposition 8

We work on the event

$$\mathcal{E}''_{\boldsymbol{W}} := \left\{\sigma_1^2(\boldsymbol{W}) \leq n\delta_W\right\} \cap \left\{c_1 \ \mathrm{tr}(\Sigma_W) \leq \frac{1}{n}\|\boldsymbol{W}\|_F^2 \leq C_1 \ \mathrm{tr}(\Sigma_W)\right\}$$

with δ_W defined in (12) and some constants $C_1 \geq c_1 > 0$, depending on γ_w. We have on the event $\mathcal{E}''_{\boldsymbol{W}}$,

$$\begin{aligned}
2\sigma_1^2(\boldsymbol{W})\frac{np}{\|\boldsymbol{W}\|_F^2} &\leq 2n\delta_W \frac{np}{\|\boldsymbol{W}\|_F^2} \\
&\leq \frac{2\delta_W}{c_1}\frac{np}{\mathrm{tr}(\Sigma_W)} && \text{by } \mathcal{E}''_{\boldsymbol{W}} \\
&= \frac{2c}{c_1}\left(\frac{np}{r_e(\Sigma_W)} + p\right) && \text{by (12)} \\
&\leq \frac{2c}{c_1}\left(\frac{n \vee p}{c'} + p\right) && \text{by } r_e(\Sigma_W) \geq c'(n \wedge p) \\
&\leq c_0(n+p) = \mu_n
\end{aligned}$$

by choosing any $c_0 \geq 2c(1 + 1/c')/c_1$. From Theorem 6 and Proposition 7 of Bing and Wegkamp (2019) with $P = \boldsymbol{I}_n$, $E = \boldsymbol{W}$ and $m = p$, we deduce

$$\widetilde{s} \leq K$$

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 $\mathcal{E}''_{\boldsymbol{W}}$,

$$\sigma_{\widetilde{s}}^2(\boldsymbol{X}) \geq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(\widetilde{s})}\|_F^2}{np - \mu_n\widetilde{s}} \geq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np}. \tag{56}$$

The first inequality uses (2.7) in Bing and Wegkamp (2019), while the second inequality uses $K \leq \bar{K}$. Further invoking (3.8) in Proposition 7 of Bing and Wegkamp (2019) yields

$$\frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np - \mu_n K} \geq \frac{\|\boldsymbol{W}\|_F^2}{np}.$$

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

$$\begin{aligned}
\mu_n \frac{\|\boldsymbol{W}\|_F^2}{np} &\geq \mu_n c_1 \frac{\mathrm{tr}(\Sigma_W)}{p} \\
&\geq 2c\left(1+\frac{1}{c'}\right)\frac{n+p}{p}\mathrm{tr}(\Sigma_W) \\
&\geq 2c\left(\mathrm{tr}(\Sigma_W) + \frac{1}{c'}\frac{n+p}{p} r_e(\Sigma_W)\|\Sigma_W\|_{\mathrm{op}}\right) \\
&\geq 2c\left(\mathrm{tr}(\Sigma_W) + (n\wedge p)\frac{n+p}{p}\|\Sigma_W\|_{\mathrm{op}}\right) && \text{by } r_e(\Sigma_W) \geq c'(n\wedge p) \\
&\geq 2c\left(\mathrm{tr}(\Sigma_W) + n\|\Sigma_W\|_{\mathrm{op}}\right) \\
&= 2n\delta_W.
\end{aligned}$$

Hence, combining all three previous displays, we derive

$$\begin{aligned}
\sigma^2_{\widetilde{s}}(\boldsymbol{X}) &\geq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np} \\
&\geq \mu_n \frac{\|\boldsymbol{W}\|_F^2}{np}\frac{np-\mu_n K}{np} \\
&\geq n\delta_W \frac{np-\mu_n K}{np} \\
&\geq \frac{1}{1+\kappa} n\delta_W && \text{by } K \leq \bar{K} \text{ and (21).}
\end{aligned}$$

Next, we prove $\sigma^2_{\widetilde{s}+1}(\boldsymbol{X}) \lesssim \delta_W$. By (2.7) in Bing and Wegkamp (2019) 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 Bing and Wegkamp (2019), 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 $\widetilde{s} \leq K$ on $\mathcal{E}''_{\boldsymbol{W}}$, we have

$$\begin{aligned}
\sigma^2_{\widetilde{s}+1}(\boldsymbol{X}) &\leq \mu_n \frac{\|\boldsymbol{X} - \boldsymbol{X}_{(K)}\|_F^2}{np-\mu_n K} \\
&\leq \mu_n \frac{np}{np-\mu_n K}\frac{\|\boldsymbol{W}\|_F^2}{np} && \text{by (3.8) of Proposition 7 in Bing and Wegkamp (2019)} \\
&\leq (1+\kappa)\mu_n \frac{\|\boldsymbol{W}\|_F^2}{np} && \text{by (21)} \\
&\leq (1+\kappa)c_0 C_1 (n+p)\frac{\mathrm{tr}(\Sigma_W)}{p} && \text{by } \mathcal{E}''_{\boldsymbol{W}} \text{ and } \mu_n = c_0(n+p) \\
&\leq \frac{(1+\kappa)c_0 C_1}{c} n\delta_W && \text{by } \mathrm{tr}(\Sigma_W) \leq p\|\Sigma_W\|_{\mathrm{op}}.
\end{aligned}$$

It remains to prove $1 - \mathbb{P}(\mathcal{E}''_{\boldsymbol{W}}) \lesssim 1/n$. First note that

$$\frac{1}{n}\|\boldsymbol{W}\|_F^2 = \sum_{j=1}^{p} \frac{1}{n}\boldsymbol{W}_{\bullet j}^\top \boldsymbol{W}_{\bullet j}.$$

By invoking Lemma 24 for fixed $j \in [p]$ and some absolute constant c, the inequality

$$\left|\frac{1}{n}\boldsymbol{W}_{\bullet j}^\top \boldsymbol{W}_{\bullet j} - [\Sigma_W]_{jj}\right| \le c\gamma_w^2 [\Sigma_W]_{jj}\sqrt{\frac{\log p}{n}}$$

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 22 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.3 Proofs for Section 4

B.3.1 Proof of Corollary 10

By Theorem 5.39 of Vershynin (2012), $\sigma_p^2(\boldsymbol{X}\Sigma_X^{-1/2}) \gtrsim n$ 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,

$$\sigma_p^2(\boldsymbol{X}) \ge \lambda_p(\Sigma_X)\sigma_p^2(\boldsymbol{X}\Sigma_X^{-1/2}) \ge \lambda_p(\Sigma_W)\sigma_p^2(\boldsymbol{X}\Sigma_X^{-1/2}) \gtrsim \lambda_p(\Sigma_W)n.$$

Corollary 10 then follows from Theorem 3 with $\widehat{\psi} = 0$, $\widehat{\eta} \gtrsim \lambda_p(\Sigma_W)$, and $\widehat{r} \le p$. ■

B.3.2 Proof of Corollary 11

Under conditions of Corollary 11, Bunea et al. (2020) proves that

$$\mathbb{P}\left\{\sigma_n^2(\boldsymbol{X}) \gtrsim \mathrm{tr}(\Sigma_W)\right\} \ge 1 - cn^{-1}.$$

We thus have $r = n$, $\widehat{\psi} = 0$, and $\widehat{\eta} \gtrsim \mathrm{tr}(\Sigma_W)/n$. Further noting that

$$\delta_W = \|\Sigma_W\|_{\mathrm{op}}\left(1 + \frac{r_e(\Sigma_W)}{n}\right) \asymp \frac{\mathrm{tr}(\Sigma_W)}{n},$$

such that $\delta_W/\widehat{\eta} \asymp 1$, we conclude

$$\begin{aligned}\mathcal{R}^*(\boldsymbol{I}_p) - \sigma^2 &\lesssim \frac{K + \log n}{n}\sigma^2 + \frac{n}{r_e(\Sigma_W)}\sigma^2 + \frac{\mathrm{tr}(\Sigma_W)}{n}\beta^\top(A^\top A)^{-1}\beta \\ &\lesssim \frac{K + \log n}{n}\sigma^2 + \frac{n}{r_e(\Sigma_W)}\sigma^2 + \frac{r_e(\Sigma_W)}{n}\|\Sigma_W\|_{\mathrm{op}}\ \beta^\top(A^\top A)^{-1}\beta.\end{aligned}$$

■

B.3.3 Proof of Theorem 12

Instead of directly applying Theorem 3, we slightly modify the proofs of Theorem 3 to obtain a sharp result for $\mathcal{R}(\widehat{A})$.

From the proof of Theorem 3, display (38) gives

$$\mathcal{R}(\widehat{A}) - \sigma^2 \le \left\|\Sigma_Z^{1/2}\left(A^\top \widehat{\alpha}_{\widehat{A}} - \beta\right)\right\|^2 + \|\Sigma_W\|_{\mathrm{op}} \left\|\widehat{\alpha}_{\widehat{A}}\right\|^2.$$

We then point out the modifications of the proof of Lemmas 17 and 18. Recall $\widehat{A} \in \mathbb{R}^{p\times \widehat{K}}$. We work on the event $\mathcal{E}^*$ defined in the proof of Theorem 3 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 $c > 0$ (Bing et al., 2020). In display (47) of Lemma 17 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

$$\begin{aligned}
\left\|\boldsymbol{X}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 &\le 2\left\|\boldsymbol{Z}AP_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 + 2\left\|\boldsymbol{W}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 \\
&\le 2\left\|\boldsymbol{Z}\Omega^{1/2}\right\|_{\mathrm{op}}^2\left\|\Sigma_Z^{1/2}(A-\widehat{A})^\top P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 \qquad (\text{by } \widehat{A}^\top\widehat{P}_{\widehat{A}}^{\perp} = 0) \\
&\quad + 4\left\|\boldsymbol{W}A^{+\top}\beta\right\|^2 + 4\left\|\boldsymbol{W}P_{\widehat{A}}A^{+\top}\beta\right\|^2 \qquad (\text{by } P_{\widehat{A}}^{\perp} = \boldsymbol{I}_p - P_{\widehat{A}}).
\end{aligned}$$

By $\mathcal{E}_{\boldsymbol{Z}}$, $\mathcal{E}'_{\boldsymbol{W}}$ and Lemma 20, after a bit algebra, we conclude

$$\begin{aligned}
\frac{1}{n}\left\|\boldsymbol{X}P_{\widehat{A}}^{\perp}A^{+\top}\beta\right\|^2 &\lesssim \left(\|A_J\|_0\frac{\log(p\vee n)}{n} + \delta_{W,J}\right)\beta^T(A^\top A)^{-1}\beta + \beta^\top A^+\Sigma_W A^{+\top}\beta \\
&\lesssim \left(\|A_J\|_0\frac{\log(p\vee n)}{n} + \|\Sigma_W\|_{\mathrm{op}}\right)\beta^T(A^\top A)^{-1}\beta + \beta^\top A^+\Sigma_W A^{+\top}\beta. \quad (57)
\end{aligned}$$

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 17, 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 18. Display (51) 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 20. For the first term of the right hand side, by (55), 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 (57) 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 23 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 $\boldsymbol{X} = \boldsymbol{Z}A^\top + \boldsymbol{W}$,

$$\begin{aligned}\left\|\boldsymbol{X}P_{\widehat{A}}^{\perp}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} &\le \left\|\boldsymbol{Z}A^\top P_{\widehat{A}}^{\perp}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} + \left\|\boldsymbol{W}P_{\widehat{A}}^{\perp}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} \\ &\le \left\|\boldsymbol{Z}(A-\widehat{A})^\top\right\|_{\mathrm{op}} + \left\|\boldsymbol{W}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} + \left\|\boldsymbol{W}P_{\widehat{A}}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} \\ &\le \left\|\boldsymbol{Z}\Omega^{1/2}\right\|_{\mathrm{op}}\sigma_1(\Sigma_Z)\left\|(A-\widehat{A})^\top\right\|_{\mathrm{op}} + \left\|\boldsymbol{W}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} + \left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\mathrm{op}}.\end{aligned}$$

By $\mathcal{E}_{\boldsymbol{Z}}$ and Lemmas 20 and 22, we have

$$\frac{1}{n}\left\|\boldsymbol{X}P_{\widehat{A}}^{\perp}A(A^\top A)^{-1/2}\right\|_{\mathrm{op}} \lesssim \delta_{W,J} + \frac{\|A_J\|_0\log(p\vee n)}{n} \lesssim \|\Sigma_W\|_{\mathrm{op}} + \frac{\|A_J\|_0\log(p\vee n)}{n}$$

with probability at least $1 - cn^{-1}$. Provided that

$$\lambda_K(A\Sigma_Z A^\top) \ge C\left(\|\Sigma_W\|_{\mathrm{op}} + \frac{\|A_J\|_0\log(p\vee n)}{n}\right)$$

for sufficiently small constant $C > 0$, we then conclude that

$$\sigma_K^2\left(\boldsymbol{X}P_{\widehat{A}}A(A^\top A)^{-1/2}\right) \gtrsim n\lambda_K(A\Sigma_Z A^\top)$$

from noting $\sigma_K^2\left(N^{1/2}(A^\top A)^{-1/2}\right) = \lambda_K(A\Sigma_Z A^\top)$. This concludes $\widehat{\eta} \gtrsim \lambda_K(A\Sigma_Z A^\top)$. 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 20 *Under conditions of Theorem 12, with probability at least* $1 - c(p\vee n)^{-1}$*, one has*

$$\frac{1}{n}\left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\mathrm{op}}^2 \lesssim \|\Sigma_W\|_{\mathrm{op}}\left(1 + \frac{\|A_J\|_{\ell_0/\ell_2}}{n}\right) := \delta_{W,J}.$$

Proof We work on the event $\widehat{K} = K$ and $\widehat{A}_I = A_I$ which holds with probability at least $1 - c(p\vee n)^{-c'}$ (Bing et al., 2020). Then

$$\begin{aligned}\left\|\boldsymbol{W}P_{\widehat{A}}\right\|_{\mathrm{op}} = \left\|\boldsymbol{W}\widehat{A}\widehat{A}^+\right\|_{\mathrm{op}} &\le \left\|\boldsymbol{W}_{\bullet I}A_I\widehat{A}^+\right\|_{\mathrm{op}} + \left\|\boldsymbol{W}_{\bullet J}\widehat{A}_J\widehat{A}^+\right\|_{\mathrm{op}} \\ &\le \left\|\boldsymbol{W}_{\bullet I}A_I(A_I^\top A_I)^{-1/2}\right\|_{\mathrm{op}}\left\|(A_I^\top A_I)^{1/2}\widehat{A}^+\right\|_{\mathrm{op}} + \|\boldsymbol{W}_{\bullet J}\|_{\mathrm{op}}\left\|\widehat{A}_J\widehat{A}^+\right\|_{\mathrm{op}}.\end{aligned}$$

Since

$$\left\|(A_I^\top A_I)^{1/2}\widehat{A}^+\right\|_{\mathrm{op}}^2 = \left\|(A_I^\top A_I)^{1/2}(\widehat{A}^\top\widehat{A})^{-1}(A_I^\top A_I)^{1/2}\right\|_{\mathrm{op}} \le 1$$

by noting $\widehat{A}^\top\widehat{A} = A_I^\top A_I + \widehat{A}_J^\top\widehat{A}_J$, and similar arguments yield

$$\left\|\widehat{A}_J\widehat{A}^+\right\|_{\mathrm{op}}^2 = \left\|\widehat{A}_J(\widehat{A}^\top\widehat{A})^{-1}\widehat{A}_J^\top\right\|_{\mathrm{op}} = \left\|(\widehat{A}^\top\widehat{A})^{-1/2}\widehat{A}_J^\top\widehat{A}_J(\widehat{A}^\top\widehat{A})^{-1/2}\right\|_{\mathrm{op}} \le 1,$$

invoking Lemma 22 to bound $\|\boldsymbol{W}_{\bullet I} A_I (A_I^\top A_I)^{-1/2}\|_{\mathrm{op}}$ and $\|W_{\bullet J}\|_{\mathrm{op}}$ gives

$$\frac{1}{n}\left\|\boldsymbol{W}_{\bullet I} A_I (A_I^\top A_I)^{-1/2}\right\|_{\mathrm{op}}^2 \lesssim \|\Psi_{II}\|_{\mathrm{op}} + \frac{\mathrm{tr}(\Psi_{II})}{n},$$

$$\frac{1}{n}\|\boldsymbol{W}_{\bullet J}\|_{\mathrm{op}}^2 \lesssim \|[\Sigma_W]_{JJ}\|_{\mathrm{op}} + \frac{\mathrm{tr}([\Sigma_W]_{JJ})}{n} \le \delta_{W,J},$$

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 $\|\Psi_{II}\|_{\mathrm{op}} \le \|[\Sigma_W]_{II}\|_{\mathrm{op}}$, $\mathrm{tr}(\Psi_{II}) \le K\|\Psi_{II}\|_{\mathrm{op}} \le K\|[\Sigma_W]_{II}\|_{\mathrm{op}}$ and $K \log n \lesssim n$. ■

B.4 Proof of Theorem 14 in Section 5

For any $\alpha \in \mathbb{R}^p$, let

$$\widehat{\mathcal{R}}(\alpha) = \frac{2}{n}\sum_{i \in D_1} [Y_i - X_i^\top \alpha]^2$$

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

$$S(\alpha) = \mathbb{E}_{(Z_*, X_*)}(Z_*^\top \beta - X_*^\top \alpha)^2, \qquad \mathcal{R}(\alpha) = S(\alpha) + \sigma^2,$$

where the expectation is over (Z_*, X_*) that are independent of α.

We have

$$\begin{aligned} S(\widehat{\alpha}) &= \mathcal{R}(\widehat{\alpha}) - \sigma^2 \\ &= (1+a)[\widehat{\mathcal{R}}(\widehat{\alpha}) - \frac{2}{n}\sum_{i \in D_1}^{n} \varepsilon_i^2] + [\mathcal{R}(\widehat{\alpha}) - (1+a)\widehat{\mathcal{R}}(\widehat{\alpha}) - (\sigma^2 - (1+a)\frac{2}{n}\sum_{i \in D_1} \varepsilon_i^2)]. \end{aligned}$$

Using $\widehat{\mathcal{R}}(\widehat{\alpha}) \le \widehat{\mathcal{R}}(\widehat{\alpha}_m)$ in the first term of the above, we have for any $m \in [M]$,

$$\begin{aligned}
S(\widehat{\alpha}) &\le (1+a)[\widehat{\mathcal{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] \\
&\quad + \max_m[\mathcal{R}(\widehat{\alpha}_m) - (1+a)\widehat{\mathcal{R}}(\widehat{\alpha}_m) - (\sigma^2 - (1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i^2)] \\
&= (1+a)[\widehat{\mathcal{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] \\
&\quad + \max_m[S(\widehat{\alpha}_m) - (1+a)\widehat{S}(\widehat{\alpha}_m) + 2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta)] \\
&\le (1+a)[\widehat{\mathcal{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2] + \max_m[S(\widehat{\alpha}_m) - (1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m)] \\
&\quad + \max_m[2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)]. \qquad (58)
\end{aligned}$$

The first term in the above can be further re-written as

$$\begin{aligned}
\widehat{\mathcal{R}}(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}^{n}\varepsilon_i^2 &= (1+a)S(\widehat{\alpha}_m) + [\widehat{\mathcal{R}}(\alpha_m) - (1+a)S(\widehat{\alpha}_m) - \frac{2}{n}\sum_{i\in D_1}\varepsilon_i^2] \\
&= (1+a)S(\widehat{\alpha}_m) + [\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m) + \frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m)] \\
&\le (1+a)S(\widehat{\alpha}_m) + \max_m[(1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m)] \\
&\quad + \max_m[\frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)].
\end{aligned}$$

Using this result in (58), we find that for any $m \in [M]$,

$$\begin{aligned}
S(\widehat{\alpha}) &\le (1+a)^2 S(\widehat{\alpha}_m) \\
&\quad + (1+a)\max_m[(1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m) - (1+a)S(\widehat{\alpha}_m)] \\
&\quad + (1+a)\max_m[\frac{4}{n}\sum_{i\in D_1}\varepsilon_i(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)] \\
&\quad + \max_m[S(\widehat{\alpha}_m) - (1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m)] \\
&\quad + \max_m[2(1+a)\frac{2}{n}\sum_{i\in D_1}\varepsilon_i(X_i^\top\widehat{\alpha}_m - Z_i^\top\beta) - \frac{a}{2}\widehat{S}(\widehat{\alpha}_m)] \\
&=: (1+a)^2 S(\widehat{\alpha}_m) + (1+a)T_1 + (1+a)T_2 + T_3 + T_4. \qquad (59)
\end{aligned}$$

Below we prove that

$$\mathbb{P}_\theta\left((1+a)T_1 + T_3 \le c_1\frac{(2+a)^3}{a}\cdot\frac{\max_m S(\widehat{\alpha}_m)\log(nM)}{n}\right) \ge 1 - c_1' n^{-1}, \qquad (60)$$

and

$$\mathbb{P}_\theta\left\{(1+a)T_2 + T_4 \le c_2\frac{(1+a)^3}{a}\sigma^2\frac{\log(nM)}{n}\right\} \ge 1 - c_2' n^{-1}, \tag{61}$$

where c_1 and c_2 depend only on $\gamma_z, \gamma_w, \gamma_\varepsilon$ from Definition 1, and $c_1, c_2 > 0$ are absolute constants. The final result follows from taking a minimum over m in (59) and combining (60) and (61) 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 (60) 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:

$$S - \widehat{S} \le \sqrt{t}\sqrt{S} \quad \Rightarrow \quad S \le (1+b)\widehat{S} + t\frac{1+b}{b}, \tag{62}$$

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 $\sqrt{S} \le \frac{1+b}{b}\sqrt{t}$, which implies $S \le \widehat{S} + t\frac{1+b}{b}$, and $\sqrt{S} > \frac{1+b}{b}\sqrt{t}$, which implies $S \le \widehat{S} + \frac{b}{1+b}S$ and thus $S \le (1+b)\widehat{S}$. Thus,

$$\begin{aligned}
\mathbb{P}_\theta\left(T_3 > t\frac{1+a/2}{a/2}\right) &\le M\max_m \mathbb{P}_\theta\left(S(\widehat{\alpha}_m) - (1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m) > t\frac{1+a/2}{a/2}\right) \\
&\le M\max_m \mathbb{P}_\theta\left(\frac{S(\widehat{\alpha}_m) - \widehat{S}(\widehat{\alpha}_m)}{\sqrt{S(\widehat{\alpha}_m)}} > \sqrt{t}\right) \qquad \text{(by (62))} \\
&\le M\max_m \mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_1}[\mathbb{E}[g_i(m)] - g_i(m)]\right| > \sqrt{t}\right), \qquad (63)
\end{aligned}$$

where we let $g_i(m) \coloneqq (Z_i^\top\beta - X_i^\top\widehat{\alpha}_m)^2/\sqrt{S(\widehat{\alpha}_m)}$ 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

$$\begin{aligned}
&\|(Z_i^\top\beta - X_i^\top\widehat{\alpha}_m)^2\|_{\psi_1} \\
&= \|Z_i^\top\beta - X_i^\top\widehat{\alpha}_m\|_{\psi_2}^2 \\
&\le \|Z_i^\top\beta - Z_i^\top A^\top\widehat{\alpha}_m\|_{\psi_2}^2 + \|W_i^\top\widehat{\alpha}_m\|_{\psi_2}^2 && \text{(since } X_i = AZ_i + W_i\text{)} \\
&= \|(\Sigma_Z^{-1/2}Z_i)^\top(\Sigma_Z^{1/2}[\beta - A^\top\widehat{\alpha}_m])\|_{\psi_2}^2 + \|(\Sigma_W^{-1/2}W)^\top(\Sigma_W^{1/2}\widehat{\alpha}_m)\|_{\psi_2}^2 \\
&= \|\Sigma_Z^{1/2}(\beta - A^\top\widehat{\alpha}_m)\|^2\|(\Sigma_Z^{-1/2}Z_i)^\top u)\|_{\psi_2}^2 && \text{(with } \|u\| = \|v\| = 1\text{)} \\
&\qquad + \|\Sigma_W^{1/2}\widehat{\alpha}_m\|^2\|(\Sigma_W^{-1/2}W)^\top v\|_{\psi_2}^2 \\
&\le c_1\|\Sigma_Z^{1/2}(\beta - A^\top\widehat{\alpha}_m)\|^2 + c_1\|\Sigma_W^{1/2}\widehat{\alpha}_m\|^2 && \text{(by Definition (1))} \\
&= c_1 S(\widehat{\alpha}_m),
\end{aligned}$$

where $c_1 = c_1(\gamma_z, \gamma_w)$. Thus,

$$\|\mathbb{E}g_i(m) - g_i(m)\|_{\psi_1} \lesssim \|g_i(m)\|_{\psi_1} \le c_1\sqrt{S(\widehat{\alpha}_m)},$$

so by Bernstein's inequality (Vershynin, 2012),

$$\mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_1}[\mathbb{E}[g_i(m)]-g_i(m)]\right|>\sqrt{t}\right)\le 2\exp\left(-n\left(\frac{t}{c_1 S(\widehat{\alpha}_m)}\wedge\sqrt{\frac{t}{c_1 S(\widehat{\alpha}_m)}}\right)\right). \quad (64)$$

Choosing $t = c_1 \max_m S(\widehat{\alpha}_m)\log(nM)/n$, and combining with (63), for $\log(M) < cn$,

$$\mathbb{P}_\theta\left(T_3 > \frac{1+a/2}{a/2}\cdot c_1\frac{\max_m S(\widehat{\alpha}_m)\log(nM)}{n}\right)\le 2/n. \quad (65)$$

We next consider T_1. For $t, b > 0$, we have

$$\widehat{S}-S\le\sqrt{t}\sqrt{S} \quad\Rightarrow\quad \widehat{S}\le\left(1+\frac{b}{1+b}\right)S + t\frac{1+b}{b}.$$

To prove this, suppose the left hand side holds and consider the cases $\sqrt{S}\le\frac{1+b}{b}\sqrt{t}$, which implies $\widehat{S}\le S+\frac{1+b}{b}t$, and $\sqrt{S}>\frac{1+b}{b}\sqrt{t}$, which implies $\widehat{S}\le[1+b/(1+b)]S$. Multiplying the right hand inequality by $(1+b)$, and choosing $b = a/2$, we find

$$\left(1+\frac{a}{2}\right)\widehat{S}-(1+a)S > t\frac{(1+a/2)^2}{a/2} \quad\Rightarrow\quad \widehat{S}-S>\sqrt{t}\sqrt{S} \quad (66)$$

Recalling

$$T_1 = \max_m[(1+\frac{a}{2})\widehat{S}(\widehat{\alpha}_m)-(1+a)S(\widehat{\alpha}_m)],$$

an application of (66) gives

$$\begin{aligned}\mathbb{P}_\theta\left(T_1 > t\frac{(1+a/2)^2}{a/2}\right) &\le M\max_m\mathbb{P}_\theta(\widehat{S}(\widehat{\alpha}_m)-S(\widehat{\alpha}_m)>\sqrt{t}\sqrt{S})\\ &\le M\max_m\mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_1}[\mathbb{E}[g_i(m)]-g_i(m)]\right|>\sqrt{t}\right)\end{aligned}$$

Choosing $t = c_1\max_m S(\widehat{\alpha}_m)\log(nM)/n$ and applying (64) with $\log(M) < cn$, we conclude

$$\mathbb{P}_\theta\left(T_1 > \frac{(1+a/2)^2}{a/2}\cdot c_1\frac{\max_m S(\widehat{\alpha}_m)\log(nM)}{n}\right)\le 2/n. \quad (67)$$

Combining (65) and (67) with a union bound and some algebra proves (60).

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 (Vershynin, 2012),

$$\mathbb{P}_\theta\left(\left|\frac{2}{n}\sum_{i\in D_2}\varepsilon_i h_i(m)\right| \ge \sqrt{t}\right) \le 2\exp(-c_2 tn/\sigma^2).$$

Choosing $t = \sigma^2\log(nM)/(c_2 n)$ completes the proof of (61). ■

Appendix C. 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 Hsu et al. (2014).

Lemma 21 *Let $\xi \in \mathbb{R}^d$ be a γ_ξ sub-Gaussian random vector. For all symmetric positive semi-definite matrices H, and all $t \ge 0$,*

$$\mathbb{P}\left\{\xi^\top H\xi > \gamma_\xi^2\left(\sqrt{\mathrm{tr}(H)} + \sqrt{2\|H\|_{\mathrm{op}}t}\right)^2\right\} \le e^{-t}.$$

Proof From Lemma 8 in Hsu et al. (2014), one has

$$\mathbb{P}\left\{\xi^\top H\xi > \gamma_\xi^2\left(\mathrm{tr}(H) + 2\sqrt{\mathrm{tr}(H^2)t} + 2\|H\|_{\mathrm{op}}t\right)\right\} \le e^{-t},$$

for all $t \ge 0$. The result then follows from $\mathrm{tr}(H^2) \le \|H\|_{\mathrm{op}}\mathrm{tr}(H)$. ■

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 Bunea et al. (2020, Theorem 10) in the sense that independence across columns of $\boldsymbol{G}$ is not required.

Lemma 22 *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,*

$$\mathbb{P}\left\{\frac{1}{n}\|\boldsymbol{G}H\boldsymbol{G}^\top\|_{\mathrm{op}} \le \gamma^2\left(\sqrt{\frac{\mathrm{tr}(H)}{n}}+\sqrt{6\|H\|_{\mathrm{op}}}\right)^2\right\} \ge 1-e^{-n}$$

Proof By definition and the property of the 1/2-net $\mathcal{N}$,

$$\|\boldsymbol{G}H\boldsymbol{G}^\top\|_{\mathrm{op}} = \sup_{u\in\mathcal{S}^{n-1}} u^\top\boldsymbol{G}H\boldsymbol{G}^\top u \le 2\sup_{u\in\mathcal{N}} u^\top\boldsymbol{G}H\boldsymbol{G}^\top u.$$

For fixed $u\in\mathcal{N}$, since $\boldsymbol{G}^\top u$ is a γ sub-Gaussian random vector, an application of Lemma 21 with $\xi=\boldsymbol{G}^\top u$, $\gamma_\xi=\gamma$ and $H=H$ yields

$$\mathbb{P}\left\{u^\top\boldsymbol{G}H\boldsymbol{G}^\top u > \gamma^2\left(\sqrt{\mathrm{tr}(H)}+\sqrt{2\|H\|_{\mathrm{op}}t}\right)^2\right\} \le e^{-t}.$$

Since $|\mathcal{N}|\le 5^n$, see Vershynin (2012, Lemma 5.2), choosing $t=3n$ and taking a union bound over $u\in\mathcal{N}$ 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 Vershynin (2012, Remark 5.40).

Lemma 23 *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 $t\ge 0$, with probability at least $1-2e^{-ct^2}$,*

$$\left\|\frac{1}{n}\boldsymbol{G}^\top\boldsymbol{G}-\Sigma_Y\right\|_{\mathrm{op}} \le \max\left\{\delta,\delta^2\right\}\|\Sigma_Y\|_{\mathrm{op}},$$

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 Bing et al. (2019) for completeness.

Lemma 24 *(Bing et al., 2019, 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*

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{t=1}^n\left(X_tY_t-\mathbb{E}[X_tY_t]\right)\right| \le \gamma_x\gamma_y t\right\} \ge 1-2\exp\left\{-c\min\left(t^2,t\right)n\right\}.$$

In particular, when $\log p\le n$, one has

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{t=1}^n\left(X_tY_t-\mathbb{E}[X_tY_t]\right)\right| \le C\sqrt{\frac{\log(p\vee n)}{n}}\right\} \ge 1-2(p\vee n)^{-c}$$

where $c\ge 2$ and $C=C(\gamma_x,\gamma_y,c)$ are some positive constants.

Appendix D. The LOVE Algorithm

For the reader's convenience, we give the specifics of estimating $\widehat{A}$ in the Essential Regression model, as developed in Bing et al. (2020). 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 1 below.

Algorithm 1 Estimate the partition of the pure variables $\mathcal{I}$ by $\widehat{\mathcal{I}}$

procedure PureVar($\widehat{\Sigma}$, δ)
  $\widehat{\mathcal{I}} \leftarrow \varnothing$.
  for all $i \in [p]$ **do**
    $\widehat{I}^{(i)} \leftarrow \left\{l \in [p] \setminus \{i\} : \max_{j \in [p]\setminus\{i\}} |\widehat{\Sigma}_{ij}| \leq |\widehat{\Sigma}_{il}| + 2\delta\right\}$
    $Pure(i) \leftarrow True$.
    for all $j \in \widehat{I}^{(i)}$ **do**
      if $\big||\widehat{\Sigma}_{ij}| - \max_{k \in [p]\setminus\{j\}} |\widehat{\Sigma}_{jk}|\big| > 2\delta$ **then**
        $Pure(i) \leftarrow False$,
        break
    if $Pure(i)$ **then**
      $\widehat{I}^{(i)} \leftarrow \widehat{I}^{(i)} \cup \{i\}$
      $\widehat{\mathcal{I}} \leftarrow$ Merge($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  return $\widehat{\mathcal{I}}$ and $\widehat{K}$ as the number of sets in $\widehat{\mathcal{I}}$

function Merge($\widehat{I}^{(i)}$, $\widehat{\mathcal{I}}$)
  for all $G \in \widehat{\mathcal{I}}$ **do** ▷ $\widehat{\mathcal{I}}$ is a collection of sets
    if $G \cap \widehat{I}^{(i)} \neq \varnothing$ **then**
      $G \leftarrow G \cap \widehat{I}^{(i)}$ ▷ Replace $G \in \widehat{\mathcal{I}}$ by $G \cap \widehat{I}^{(i)}$
      return $\widehat{\mathcal{I}}$
  $\widehat{I}^{(i)} \in \widehat{\mathcal{I}}$ ▷ add $\widehat{I}^{(i)}$ in $\widehat{\mathcal{I}}$
  return $\widehat{\mathcal{I}}$

Given estimates $\widehat{K}$ and $\widehat{\mathcal{I}}$ as outputs of Algorithm 1, we compute, for each $a \in [\widehat{K}]$ and $b \in [\widehat{K}] \setminus \{a\}$,

$$\left[\widehat{\Sigma}_Z\right]_{aa} = \frac{1}{|\widehat{I}_a|(|\widehat{I}_a| - 1)} \sum_{i,j \in \widehat{I}_a, i \neq j} |\widehat{\Sigma}_{ij}|, \quad \left[\widehat{\Sigma}_Z\right]_{ab} = \frac{1}{|\widehat{I}_a||\widehat{I}_b|} \sum_{i \in \widehat{I}_a, j \in \widehat{I}_b} \widehat{A}_{ia}\widehat{A}_{ib}\widehat{\Sigma}_{ij}, \tag{68}$$

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 $\widehat{I}_k$,

$$\text{Pick an element } i \in \widehat{I}_k \text{ at random, and set } \widehat{A}_{i\cdot} = e_k; \tag{69}$$

$$\text{For the remaining } j \in \widehat{I}_k \setminus \{i\}, \text{ set } \widehat{A}_{j\cdot} = \text{sign}(\widehat{\Sigma}_{ij}) \cdot e_k. \tag{70}$$

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 Bing et al. (2020) given by

$$\widehat{A}_{j\cdot} = \arg\min_{\beta^j} \left\{ \|\beta^j\|_1 : \left\| \widehat{\Sigma}_Z \beta^j - (\widehat{A}_{\widehat{I}}^\top \widehat{A}_{\widehat{I}})^{-1} \widehat{A}_{\widehat{I}}^\top \widehat{\Sigma}_{\widehat{I}j} \right\|_\infty \le \mu \right\} \tag{71}$$

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 $1 \le q \le \infty$ (Bing et al., 2020, Theorem 5).

Appendix E. 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 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 $\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{U}}_K$ where $\widetilde{\boldsymbol{U}}_K$ is the matrix of the first K right singular vectors of $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_{1\bullet}^\top, \ldots, \widetilde{\boldsymbol{X}}_{n\bullet}^\top)^\top$. Bai (2003) shows that

$$V_t^{-1/2} \left(\widehat{\boldsymbol{Y}}_t - \boldsymbol{Z}_{t\bullet}^\top \beta \right) \to N(0,1), \qquad \text{for any } 1 \le t \le n \tag{72}$$

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 Fan et al. (2013). 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 6 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 (29), The uniform consistency in Fan et al. (2013) additionally requires $n = o(p^2)$. As a result, all previous results are asymptotic statements as $n, p \to \infty$.

By contrast, our Corollaries 5, 6 and 9 are non-asymptotic statements which hold for any finite K, n and p. Moreover, they only requires the sub-Gaussian tail assumptions in Definition 1 and $K \log n \lesssim n$. As detailed in Section 3.2, our conditions on the signal $\lambda_K(A\Sigma_Z A^\top)$ are much weaker than (29) to derive the risk of PCR-K.

Under condition (29), as assumed in the aforementioned literature, the prediction risk in our Corollary 6 reduces to

$$\mathcal{R}(\boldsymbol{U}_K) - \sigma^2 = O_p\left(\frac{\sigma^2}{n} + \frac{\|\Sigma_W\|_{\mathrm{op}}}{p} + \frac{\|\Sigma_W\|_{\mathrm{op}}}{n} \right).$$

This rate coincides with that of V_t, introduced in (72). Under conditions in Fan et al. (2013), 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.

References

Seung C. Ahn and Alex R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Jushan Bai and Serena Ng. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150, 2006. doi: 10.1111/j.1468-0262.2006.00696.x.

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.

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.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *In arXiv:1906.11300*, 2019.

Mikhail Belkin, Daniel Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *In arXiv:1806.05161*, 2018a.

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

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

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

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

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. doi: 10.1214/18-AOS1774. URL `https://doi.org/10.1214/18-AOS1774`.

Xin Bing, Florentina Bunea, and Marten Wegkamp. Inference in interpretable latent factor regression models. *In arXiv:1905.12696*, 2019.

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.

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. doi: 10.3150/14-BEJ602. URL `https://doi.org/10.3150/14-BEJ602`.

Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Interpolation under latent factor regression models. *In arXiv:2002.02525*, 2020.

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.

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

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

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

Harold Hotelling. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79, 1957.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, June 2014. ISSN 1615-3375. doi: 10.1007/s10208-014-9192-1.

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. ISSN 0304-4076. High Dimensional Problems in Econometrics.

Maurice G. Kendall. *A course in multivariate analysis*. Hafner Pub. Co., 1957.

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.

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

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.

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

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *In arXiv:2005.08054*, 2020.

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. ISSN 01621459.

Roman Vershynin. *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 2012.

Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1):252–273, 2003.