Essential Regression

Xin Bing∗ Florentina Bunea† Marten Wegkamp‡ Seth Strimas-Mackey§

Abstract

Essential Regression is a new type of latent factor regression model, where unobserved factors $Z \in \mathbb{R}^K$ influence linearly both the response $Y \in \mathbb{R}$ and the covariates $X \in \mathbb{R}^p$ with $K \ll p$. Its novelty consists in the conditions that give $Z$ interpretable meaning and render the regression coefficients $\beta \in \mathbb{R}^K$ relating $Y$ to $Z$ – along with other important parameters of the model – identifiable. It provides tools for high dimensional regression modelling that are especially powerful when the relationship between a response and essential representatives $Z$ of the $X$-variables is of interest.

Since in classical factor regression models $Z$ is often not identifiable, nor practically interpretable, inference for $\beta$ is not of direct interest and has received little attention. We bridge this gap in E-Regressions models: we develop a computationally efficient estimator of $\beta$, show that it is minimax-rate optimal (in Euclidean norm) and component-wise asymptotically normal, with small asymptotic variance. Inference in Essential Regression is performed after consistently estimating the unknown dimension $K$, and all the $K$ subsets of the $X$-variables that explain, respectively, the individual components of $Z$. It is valid uniformly in $\beta \in \mathbb{R}^K$, in contrast with existing results on inference in sparse regression after consistent support recovery, which are not valid for regression coefficients of $Y$ on $X$ near zero.

Prediction of $Y$ from $X$ under Essential Regression complements, in a low signal-to-noise ratio regime, the battery of methods developed for prediction under different factor regression model specifications. Similarly to other methods, it is particularly powerful when $p$ is large, with further refinements made possible by our model specifications.

Moreover, Essential Regression provides a statistical platform for analysis in regression with clustered covariates, with or without overlap. This allows us to address possible inferential questions in post clustering-inference, and subsequently provide guidelines regarding the use and misuse of cluster averages as popular dimension reduction devices in high-dimensional regression.

We benchmark the Essential Regression methodology on a recently collected data set for the study of the effectiveness of a new SIV vaccine. Our analysis enables the determination of the most important antibody-centric mechanisms associated with the vaccine response.

Keywords: High dimensional regression, latent factor model, identification, uniform inference, minimax estimation, pure variables, post clustering inference/regression, adaptive estimation

∗Department of Statistical Science, Cornell University, Ithaca, NY. E-mail: xb43@cornell.edu.
†Department of Statistical Science, Cornell University, Ithaca, NY. E-mail: fb238@cornell.edu.
‡Department of Mathematics and Department of Statistical Science, Cornell University, Ithaca, NY. E-mail: mhw73@cornell.edu.
§Department of Statistical Science, Cornell University, Ithaca, NY. E-mail: scs324@cornell.edu.
1 Introduction

We introduce the Essential Regression (E-Regression) model, as an alternative to the ubiquitous sparse high-dimensional linear regression on \( p \) variables. It is a class of regression models tailored to applications where the relation between the dependent variable \( Y \) and representatives of groups of components of the independent variables \( X \), rather than between \( Y \) and the components of \( X \), is of main interest. A specific challenge addressed within the E-Regression framework is the definition of representatives in a mathematically coherent and practically interpretable way.

Formally, E-Regression is a new variant of the more classical factor regression model, introduced by Stock and Watson (2002a,b), which postulates the existence of an unobserved, zero mean, random vector \( Z \in \mathbb{R}^K \), for some unknown \( K < p \), that is connected to the observed pair \((X,Y) \in \mathbb{R}^p \times \mathbb{R} \) via the model

\[
Y = Z^T \beta + \varepsilon \\
X = AZ + W.
\]

The dimension \( K \), matrix \( A \in \mathbb{R}^{p \times K} \) and vector \( \beta \in \mathbb{R}^K \) are unknown, and \( Z, \varepsilon \) and \( W \) are independent. Furthermore, \( \varepsilon \) and \( W \) have zero mean, and unknown variance \( \sigma^2 \) and diagonal covariance matrix \( \Sigma^w \), respectively. In contrast to sparse regression, where only few components of the observable \( X \) are assumed to directly influence \( Y \), our framework allows for all \( p \)-components of \( X \) to influence \( Y \), but mediated through the lower dimensional random vector \( Z \). The mediator \( Z \) is not observed, and made interpretable via a modeling assumption through which each component of \( Z \) is given the physical meaning of a small group of the \( X \)-variables.

Factor regression models, and their many variants (Stock and Watson, 2002a,b; Bai and Ng, 2006; Bair et al., 2006; Boivin and Ng, 2006; Blei and McAuliffe, 2007; Bai and Ng, 2008; Hahn et al., 2013; Kelly and Pruitt, 2015; Fan et al., 2017) have been introduced to improve the prediction of \( Y \in \mathbb{R} \) from \( X \in \mathbb{R}^p \), when \( p \) is very large and the components of \( X \) are highly correlated. For this purpose, the matrix \( A \) in (2) need only be unique up to generic invertible matrix transformations. This no longer suffices for the primary goal of this work, inference on the lower dimensional vector \( \beta \), when two other aspects become important:

- \( Z \) must be interpretable so that regression model (1) is interpretable;
- \( \beta \) must be uniquely defined.

Both desiderata are met by placing the following assumptions on \( A \) and the covariance matrix \( \Sigma^z \) of \( Z \), which differ from popular assumptions in the factor analysis literature, such as, for instance, Assumption 5 in Section 3.1.

Assumption 1.

\((A0)\) \( \|A_j\|_1 \leq 1 \) for all \( j \in [p] \).

\((A1)\) For every \( k \in [K] \), there exists at least two \( j \neq \ell \in [p] \), such that \( |A_{jk}| = |A_{\ell k}| = 1 \) and \( A_{jk'} = A_{\ell k'} = 0 \) for any \( k' \neq k \).

\((A2)\) The covariance matrix \( \Sigma^z := \text{Cov}(Z) \) is positive definite. There exists a constant \( \nu > 0 \) such that

\[
\min_{1 \leq a < b \leq K} \left( \Sigma^z_{aa} \wedge \Sigma^z_{bb} - |\Sigma^z_{ab}| \right) > \nu.
\]
Assumption 1, first introduced in Bing et al. (2017), guarantees that $A$ and $\Sigma^z$ are identifiable, up to signed permutations. We refer to this work for an in-depth comparison with the rich literature on factor models of type (2) and detailed explanation of assumptions $(A0) – (A2)$, and only offer a short summary here. The first assumption is a scale assumption, and helps interpret $A$ as a membership matrix, for the applications to clustering that we consider below in Sections 4 and 6. It further allows for $X_j = E_j$ for some components $X_j$ of $X$ that are not related to $Z$. The third assumption allows us to depart from the widely used, and restrictive, assumption of independence among the latent factors, and it implies the minimal requirement that the factors must be different. The second assumption, called the pure variable assumption, is crucial for our model interpretation. It states that every $Z_k$, $k \in [K]$, must have at least two components of $X$, the pure variables, solely associated with it, up to additive noise with possibly different variance levels. These pure variables impart on each $Z_k$ their meaning and the regression model (1) becomes interpretable.

An illustration of our model specifications is given in Figure 1 below. Three latent factors $(Z_1, Z_2, Z_3)$ are connected to $(X_1, \ldots, X_{10})$. The numbers on the edges denote the strength of this connection quantified by the corresponding entries in $A$. $X_1$ and $X_2$ are only connected to $Z_1$, and as such, are pure variables.

Figure 1: An illustrative example of Essential Regression

A similar, data-driven, figure is presented in Section 6, in which we show that the E-Regression model fits the data collected in a new SIV-study (SIV is the non-human primate equivalent of HIV), and offers important insights into mechanisms driving the vaccine response. E-Regression provides a novel and scientifically-desirable way of modeling a response $Y$ directly at the mechanism ($Z$) level, a task that is difficult to accomplish via a traditional regression of $Y$ on $X$.

To the best of our knowledge, factor regression models under Assumption 1 have not been studied. However, factor models of type (2), under assumptions similar in spirit to ours have appeared as early as the 1950’s in the statistical literature. An instance is Anderson and Rubin.
(1956), who provide practical interpretations of the pure variable assumption, but only consider the situation when both the pure variable set and $K$ are known. Related models of type (2) have been more recently revived in the theoretical computer science literature (Arora et al., 2012, 2013; Bittorf et al., 2012). These works make use of the pure variable assumption, do not require their set be known, but do require that $K$ is a priori known. While this is a relaxation of the assumptions proposed by Anderson and Rubin (1956), it has only been studied in the context of the specific class of topic models, when the matrices $X$, $A$ and $Z$ all have positive and appropriately scaled entries. We refer again to (Bing et al., 2017, Section 4.4) for an in-depth comparison with this, and other related literature on factor models (2).

The study of factor regression models satisfying (1) and (2), under Assumption 1, brings as a novel element, relative to classical factor regression models, not only the fact that the assumption leads to parameter identifiability and interpretability of $Z$, but also that it can be used constructively for estimation in this model, as detailed in the following section.

1.1 Our contributions

We begin by summarizing the model parameters, the nature of the data, as well as the relation between parameter dimensions and sample size. Throughout this work we assume that we have access to an i.i.d. sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ of $(X,Y)\in\mathbb{R}^p \times \mathbb{R}$.

We denote by $I \subseteq \{1,\ldots,p\}$ the index set of the pure $X$-variables. The following quantities are unknown and will be estimated from the data, under the Essential Regression model: $A$, $I$, $K$, $\beta$, $\sigma^2$, $\Sigma^w$, $\Sigma^z$. We allow for $p>n$, while $K<p$. In this work, we consider the case of non-sparse $\beta$, and $K<n$, but allow $K$ to grow with the sample size $n$. The complementary cases of $K>n$ and $\beta$ sparse will be studied in a follow-up work.

1.1.1 Estimation and inference for $\beta$

In any factor regression model the regression coefficient vector is given by the expression

$$\beta = (\Sigma^z)^{-1}(A^T A)^{-1} A^T \text{Cov}(X,Y).$$

(3)

This representation is not unique, if $A$ and $\Sigma^z$ are not uniquely defined. For independent latent factors $Z$, (Anderson and Rubin, 1956) provide assumptions similar to Assumption 5 reviewed in Section 3.1 below under which $A$ is uniquely defined. They refer to these assumptions as mathematically convenient, and acknowledge that they do no lead to a practically interpretable model. Indeed, as shown in Lemma 9 in Section 3.1 below, under the classical factor model and factor regression model identifiability assumptions, the latent factors $Z$ mimic the principal components of $X$, making them difficult to interpret as covariates in (1).

To the best of our knowledge, the asymptotic distribution of estimators of $\beta$ from an i.i.d. sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ in classical factor regression models has not been investigated, primarily because when $Z$ lacks interpretability, so does $\beta$. Moreover, the rate-optimality of estimators of $\beta$ in factor regression, and estimation procedures that adapt to it, have also not been studied.

We bridge this gap, under the Essential Regression framework. The pure variable assumption $A I$ is at the heart of our solution, and is the reason $Z$ can be interpreted to begin with. It also allows us to show that $\beta$ is uniquely defined, making inferential questions about $\beta$ well-posed.

We show in Section 2.1 that the generic expression (3) of $\beta$ simplifies to

$$\beta = (\Sigma^z)^{-1}(A_I^T A_I)^{-1} A_I^T \text{Cov}(X_I,Y).$$

(4)
in our Essential Regression framework. Here \( A_I \) is the sub-matrix of \( A \) with row indices corresponding to indices in the pure variable set \( I \).

Under (A1) of Assumption 1, \( \Sigma = \text{Cov}(X) \) contains a sub-matrix with off-diagonal block structure, and averages of the block entries equal the entries of \( \Sigma_z \). The position of these blocks within \( \Sigma \) is given by the partition of the index set \( I \subseteq [p] \) of pure variables. We showed in Bing et al. (2017) that the set of pure variables \( I \) and its partition can be uniquely reconstructed from \( \Sigma \). Therefore, \( \Sigma_z \) and \( A_I \) are uniquely defined, and so is \( \beta \). The details of this derivation are given in Section 2.1. In Section 2.2 we use the representation (4) and plug-in estimators of the unknown quantities to construct our proposed estimator \( \hat{\beta} \), analyzed in Sections 2.5 and 2.6. We employ the LOVE algorithm developed in Bing et al. (2017) to estimate \( I \), its partition, and \( K \).

To benchmark the quality of estimation of \( \beta \), under the Essential Regression framework, we prove in Theorem 3 of Section 2.4 that the minimax optimal rate of estimating \( \beta \) in the \( \ell_2 \)-norm in \( \mathbb{R}^K \) is \((1 \lor \|\beta\|/\sqrt{m})\sqrt{K/n}\) in our model with \( K < n \). The quantity \( m \) is the size of the smallest group of pure variables. The factor \( \sqrt{K/n} \) is the standard minimax rate of estimation in linear regression with observed \( Z \), and sub-Gaussian errors. The factor multiplying it can be viewed as the price to pay for not observing \( Z \). It quantifies the trade-off between not observing \( Z \), with strength \( \|\beta\| \), the \( \ell_2 \) norm of \( \beta \), and the number of times, \( m \), each component of \( Z \) is partially observed, up to additive error. The ratio \( \|\beta\|/\sqrt{m} \) indicates that, under the Essential Regression framework, the fact that \( Z \) is not observed is alleviated by the existence of pure variables, and the quality of estimation is expected to increase as \( m \) increases. Our minimax results are new in the literature on factor regression models. In Section 2.4 we explain how they connect to recent results obtained in errors-in-variables models Belloni et al. (2017), which correspond to \( K = p \), and \( A = I_p \).

We show in Theorem 4 of Section 2.5 that the proposed estimator \( \hat{\beta} \) is minimax rate optimal, up to logarithmic factors in \( n \) and \( p \). Our result uses the fact that only the estimation of the sub-matrix \( A_I \), instead of the entire \( p \times K \) matrix \( A \), is involved in the construction of \( \hat{\beta} \). In Section 2.5 we present a detailed rate comparison with other natural competitors, including an estimator of \( \beta \) that utilizes estimators of the full matrix \( A \).

In Section 2.6, we show that our estimator \( \hat{\beta} \) is component-wise asymptotically normal. Its asymptotic variance agrees in order with the information bound in our Essential Regression model and can be consistently estimated (Proposition 7). The analysis of \( \hat{\beta} \) relies on being able to consistently identify the pure variables. This is done by using the sample \( X_1, \ldots, X_n \) alone, without using \( Y_1, \ldots, Y_n \), and consequently, inference for \( \beta \), at the coarser resolution level provided by the essence \( Z \), is valid uniformly over \( \beta \). This is in contrast with inference in direct sparse regression of \( Y \) on \( X \), after consistently estimating the support of \( \beta \), which is valid only for regression coefficients above the minimax optimal \( O(\sqrt{\log p/n}) \) level (Bunea, 2008; Bühlmann and van de Geer, 2011; Giraud, 2015).

### 1.1.2 Prediction of \( Y \) from \( X \) via Essential Regression

In general factor regression (FR) models (1) – (2), at the population level, the best linear predictor of \( Y \) takes the form

\[
Y_{FR}^* = X^T A [\text{Cov}(A^T X)]^{-1} \text{Cov}(A^T X, Y).
\]

It is uniquely defined for any matrix \( A \), modulo invertible transformations \( Q \). Since the matrix \( A \) is identifiable up to a trivial orthogonal transformation (a signed permutation matrix) in our
Essential Regression model, we use the above expression, combined with a plug-in estimate of $A$, to construct in-sample predictors $\hat{Y}$ of the observed data vector $Y \in \mathbb{R}^n$. Details are given in Section 3, where we state in Theorem 8 the in-sample prediction risk bound

$$\frac{1}{n} \mathbb{E} \left[ \| \hat{Y} - Z\beta \|_2^2 \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\| \beta \|_2^2 \Lambda_{\min}}{1 + s_J \log (p \vee n)} \right \}$$

(6)

In Theorem 8, we additionally prove an almost identical bound for the risk of predictors $\hat{Y}_\ast$ of the response $Y_\ast \in \mathbb{R}$ for a single new data vector $X_\ast \in \mathbb{R}^p$. The quantity $\Lambda_{\min}$ in (6) denotes the smallest eigenvalue of $A^T A$ and $s_J$ is the size of the support of the sub-matrix $A_J$ of $A$, corresponding to the set $J \subseteq [p]$ of non-pure variables. The first term in the above rate $K\sigma^2/n$ reflects the dimension reduction (from $p$ to $K$) and the second term is the price to pay for not observing $Z$. Once again, we observe the trade-off between the strength of the non-observed signal $Z$, quantified by $\| \beta \|_2^2$, and the amount of information given by partially observing it, quantified by $\Lambda_{\min}$. Furthermore, when non-pure variables are associated with fewer than $K$ latent factors, the prediction risk becomes smaller, as in that case $A_J$ is sparse and $s_J$ is small. The sparsity of $A_J$ is also a desired feature for the applications of E-Regression to regression with clustered variables described in Section 1.1.3 below. Our illustration, Figure 1, reflects this situation as the non-pure variables $X_3, X_4, X_7, X_8$ are associated to only two of the three latent factors. More generally, the sparsity of $A_J$ is supported by the data analysis presented in Section 6.

In Section 3.1 we contrast our model-based prediction strategy with Principal Component Regression (PCR), a popular model-free method that predicts $Y$ from the first $K$ principal components of $X$. PCR dates several decades back to Kendall (1957); Hotelling (1957), with more recent overviews in Cook (2007); Izenman (2008). At the population level, the predictor is given by

$$Y_{\ast PCR} := X^T U_K [\text{Cov}(U_K^T X)]^{-1} \text{Cov}(U_K^T X, Y),$$

(7)

an expression analogous to (5), when $A$ is replaced by the matrix $U_K$ that has the first $K$ eigenvectors of $\Sigma$ as its columns.

We establish the population level asymptotic equivalence between the predictors (5) and (7) in Lemma 9 in Section 3.1. This result shows, at the population level, that any factor regression model provides a platform that justifies this popular dimension reduction scheme, provided either the variances in $\Sigma^w$ are all equal, or a suitable condition based on $\lambda_K(A\Sigma^z A^T)/\lambda_1(\Sigma^w)$ is placed. The repercussions of this fact have been exploited in the classical factor regression literature, in which sample based prediction is given by $\hat{Y}_{PCR}$ corresponding to estimating the principal components in $Y_{PCR}$. A short list of references of works devoted to analyzing the quality of this predictor includes Stock and Watson (2002a,b), who gave sufficient assumptions on $A$, $\Sigma^z$ and $\Sigma^w$ that allowed for consistent prediction of $Y$, with further results established by Boivin and Ng (2006); Bair et al. (2006); Bai and Ng (2008, 2006); Fan et al. (2017). In Section 3.1 we review the assumptions under which PCR-based prediction is consistent, and show that under structural model assumptions, such as those provided by Essential Regression, a model-based predictor is consistent under weaker assumptions. Moreover, to the best of our knowledge, our analysis is the first to allow the number of factors $K$ to increase with $n$, and to state a finite sample risk upper bound that reveals the interplay among the quantities that govern prediction accuracy in this regime.
1.1.3 Essential Regression as Regression with Clustered Predictors

The purpose of Section 4 is to show that E-Regression can be used as a vehicle for model-based clustering and subsequent regression on cluster-related quantities. A routinely used strategy in practical applications that involve a very large number of covariates, is first to employ an algorithm that returns clusters of the \(X\)-variables, typically non-overlapping, then to regress a response vector \(Y\) on cluster averages, and finally to perform analysis on the resulting model (Bühlmann et al., 2013). Oftentimes it is unclear how the response \(Y\) depends on clusters of \(X\) that are determined independently of \(Y\). More generally, an approach for conducting post-clustering inference in regression is under-explored, as the few existing results focus on prediction. With Essential Regression we provide a model-based framework for such analyses.

Clusters of the \(X\)-variables are obtained relative to the support of the columns of \(A\), with index sets given by

\[
G_k := \{j \in [p] : |A_{jk}| > 0\}, \quad \text{for } k \in [K],
\]

and are allowed to overlap. This program, regarding the clustering of \(X\), without the regression step, has been carried out in Bing et al. (2017). They showed that clusters can be estimated consistently under weak signal strength conditions.

Within our E-Regression framework, we distinguish between two post-clustering problems: inference and prediction. We can interpret the matrix \(A\) as a cluster allocation matrix and the inference carried out at the level of the latent factors \(Z\), as inference carried out at the level of the cluster centers, but caution against replacing components of \(Z\) by cluster averages. Indeed, we prove in Section 4 that replacing \(Z\) by the weighted averages

\[
\bar{X} := (A^T A)^{-1} A^T X,
\]

and subsequently regressing on \(\bar{X}\), would not estimate \(\beta\). However, this can be immediately corrected by regression on predictors \(\tilde{Z}\) of \(Z\), obtained from appropriate cluster averages, exercising care when clusters overlap. With this correction, we obtain exactly the estimator of \(\beta\) analyzed in Section 2.5, and we can interpret the inferential tools developed in Section 2.6 as tools for post-clustering inference in regression. Prediction of \(Y\) requires less care as the cluster (weighted) averages \(\bar{X}\) in (8) has the same prediction error as that performed from \(\tilde{Z}\). The resulting predictor corresponds to the one already analyzed in Section 3 and our model formally justifies prediction from cluster averages. The prediction error adapts to the unknown reduced dimension \(K\), and becomes smaller not only when \(n\) increases, but typically also when \(p\) grows, in agreement with all prior results on prediction via factor regression models, irrespective of their specifications. Moreover, prediction with clustered variables, whenever appropriate, provides an alternative to prediction via sparse regression in high dimensions, with differences particularly pronounced whenever the level of sparsity is not high and when the multi-collinearity among the \(X\)-variables is strong, as further verified empirically in Section 5.2, for a general assignment matrix \(A\) that allows for cluster overlap.

1.2 Organization of the paper

Identifiability of \(\beta\) is studied in Section 2.1. The estimation procedure of \(\beta\) is proposed in Section 2.2. We state our main assumptions, in addition to model specification Assumption 1, in Section 2.3. Sections 2.4 and 2.5 provide the lower bound and upper bound of the \(\ell_2\)-norm convergence rate of \(\hat{\beta}\), respectively. The asymptotic normality of \(\hat{\beta}\) is studied in Section 2.6. In Section 3, we
propose our prediction procedure and provide the rate of its in-sample and new-data prediction risks. Section 3.1 contains detailed comparison with factor regression model. The application of Essential Regression to clustered predictors is discussed in Section 4. The simulation result is included in Section 5, followed by a real data analysis in Section 6. All proofs are deferred to the supplement.

1.3 Notation

For any positive integer \( q \), we let \([q] = \{1, 2, \ldots, q\}\). For two numbers \( a \) and \( b \), we write \( a \lor b := \max\{a, b\} \) and \( a \land b := \min\{a, b\} \). For a set \( S \), we use \( |S| \) to denote its cardinality. For a generic vector \( v \), we let \( \|v\|_q = (\sum_i |v_i|^q)^{1/q} \) denote its \( \ell_q \) norm for \( 1 \leq q \leq \infty \) with the convention that \( s^{1/q} = 1 \) for \( q = \infty \). In particular, we write \( \|v\| = \|v\|_2 \). We also write \( \|v\|_0 = |\text{supp}(v)| \). Let \( Q \) be any matrix. We use \( \|Q\|_{op} = \sup_{v \in S^{d-1}} \|Qv\| \), \( \|Q\| = (\sum_{i,j} Q_{ij}^2)^{1/2} \) and \( \|Q\|_\infty = \max_{i,j} |Q_{ij}| \) for its operator norm, Frobenius norm and element-wise maximum norm, respectively. We write \( \|Q\|_0 \) for its vectorized \( \ell_0 \) norm. For a symmetric matrix \( Q \in \mathbb{R}^{d \times d} \), we denote by \( \lambda_k(Q) \) its \( k \)th largest eigenvalue for \( k \in [d] \). For a positive semi-definite symmetric matrix, we will frequently use the fact that \( \lambda_1(Q) = \|Q\|_{op} \).

We use \( \mathcal{H}_d \) to denote the set of all \( d \times d \) signed permutation matrices and \( S^{d-1} \) to represent the space of the unit vectors in \( \mathbb{R}^d \). We denote by \( I_d \) the \( d \times d \) identity matrix, by \( 1_d \) the \( d \)-dimensional vector with entries equal to 1 and by \( \{e_j\}_{1 \leq j \leq d} \) the canonical basis in \( \mathbb{R}^d \). For any two sequences \( a_n \) and \( b_n \), \( a_n \preceq b_n \) (or \( a_n = O(b_n) \)) stands for there exists constant \( C > 0 \) such that \( a_n \leq Cb_n \). We write \( a_n \asymp b_n \) if \( a_n \preceq b_n \) and \( b_n \preceq a_n \). We also use \( a_n = o(b_n) \) to denote \( a_n/b_n \to 0 \) as \( n \to \infty \).

Based on (A1) of Assumption 1, we denote the set of pure variables as

\[
I = \bigcup_{k=1}^K I_k, \quad I_k = \{i \in [p] : |A_{ik}| = 1, A_{ik'} = 0, \text{ for any } k' \neq k\}.
\]

(9)

Its complement set is called the non-pure variable set \( J := [p] \setminus I \). We then partition the \( p \times K \) matrix \( A \) as \( A_I \in \mathbb{R}^{|I| \times K} \) and \( A_J \in \mathbb{R}^{|J| \times K} \) corresponding to \( I \) and \( J \), respectively. Finally, we let \( m := \min_k |I_k| \) denote the size of the smallest group of pure variables.

2 Estimation and inference in Essential Regression

2.1 Identifiability of \( \beta \)

Under model (1), we have \( Y = Z^T \beta + \varepsilon \), and thus the coefficient \( \beta \) satisfies

\[
\beta = [\text{Cov}(Z)]^{-1} \text{Cov}(Z, Y) = (\Sigma^z)^{-1} \text{Cov}(Z, Y).
\]

(10)

Since under model (2) we also have \( X = AZ + W \), which implies \( X_I = A_I Z + W_I \), then

\[
\text{Cov}(Z, Y) = (A_I^T A_I)^{-1} A_I^T \text{Cov}(X_I, Y),
\]

as under our assumptions \( \text{rank}(A_I) = K \), leading to the following equivalent expressions for \( \beta \):

\[
\beta = (\Sigma^z)^{-1} (A_I^T A_I)^{-1} A_I^T \text{Cov}(X, Y)
\]

(11)

\[
= (\Sigma^z)^{-1} (A_I^T A_I)^{-1} A_I^T \text{Cov}(X_I, Y).
\]

(12)
Theorem 1 in Bing et al. (2017) shows that, under Assumption 1, the matrices \(A\) and \(\Sigma_z\) can be uniquely determined from \(\Sigma\), up to a signed permutation matrix \(P\). As a result, \(\beta\) can also be recovered from (12) up to \(P^T\), as summarized in the proposition below. The proof can be found in the supplement.

**Proposition 1.** The quantities \(\Sigma\) and \(\text{Cov}(X,Y)\) define \(\beta\) uniquely, via (12), up to the permutation matrix \(P^T\).

The permutation matrix \(P\) will not affect either inference or the prediction of \(Y\). Indeed, writing \(\tilde{A} = AP\), \(\tilde{Z} = P^T Z\) and \(\tilde{\beta} = P^T \beta\), one still has \(Y = \tilde{Z}^T \tilde{\beta} + \varepsilon\) and \(X = \tilde{A} \tilde{Z} + W\).

### 2.2 Estimation of \(\beta\)

We assume that the data consists of \(n\) independent observations \((X_1,Y_1),\ldots,(X_n,Y_n)\) of \((X,Y)\) that satisfies (1) and (2). Let \(\hat{\Sigma} = n^{-1} \sum_{i=1}^{n} X_i X_i^T\) denote the sample covariance matrix. Motivated by equation (12), we consider a plug-in estimator of \(\beta\) via the following steps:

1. Obtain estimates \(\hat{K}, \{\hat{I}_1,\ldots,\hat{I}_K\}, \hat{A}_{\hat{I}}\) and \(\hat{\Sigma}_z\) from \(\hat{\Sigma}\) with tuning parameter \(\delta\) by using Algorithm 1 in Bing et al. (2017). For the reader’s convenience, the procedure is stated in Appendix A of 6.

2. Compute
   \[
   \hat{h} = \frac{1}{n} \left(\hat{A}_{\hat{I}}^T \hat{A}_{\hat{I}}\right)^{-1} \hat{A}_{\hat{I}}^T X_{\hat{I}}^T Y. \tag{13}
   \]

Provided \(\hat{\Sigma}_z\) is non-singular, estimate \(\beta\) by
   \[
   \hat{\beta} = \left(\hat{\Sigma}_z\right)^{-1} \hat{h}. \tag{14}
   \]

The procedure requires a single tuning parameter \(\delta\), and that \(K < n\). We show in Section 4 that \(\hat{\beta}\) coincides with the ordinary least squares estimator that minimizes \(\|Y - \hat{Z}_{(\hat{I})} \beta\|^2\) over \(\beta\), based on an appropriately constructed predictor \(\hat{Z}_{(\hat{I})}\) of \(Z\). We further discuss other possible estimators, including the cases \(K > n\) and \(\hat{\Sigma}_z\) singular, in Section 2.5, and contrast their rate performance with that of our recommended estimator.

### 2.3 Assumptions

The implementation of the above procedure is free of distributional assumptions. We evaluate it under three additional assumptions. First, we make the following distributional specifications for \(\varepsilon, W\) and \(Z\) defined in model (1) – (2):

**Assumption 2.** Let \(\gamma_{\varepsilon}, \gamma_{W}, \gamma_z\) and \(B_z\) be positive finite constants. Assume \(\varepsilon\) is \(\gamma_{\varepsilon}\)-subgaussian\(^1\) and \(W\) has independent \(\gamma_{W}\)-subgaussian entries. Further assume \(\|\Sigma_z\|_\infty \leq B_z\) and the random vector \((\Sigma_z)^{-1/2} Z\) is \(\gamma_z\)-subgaussian\(^2\).

---

\(^1\)A mean zero random variable \(x\) is called \(\gamma\)-subgaussian 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 \(\gamma\)-subgaussian if \(\langle x, v \rangle\) is \(\gamma\)-subgaussian for any unit vector \(v\).
The quality of our estimator $\hat{\beta}$ depends on how well we estimate $K$, $I$, its partition $\{I_k\}_{1 \leq k \leq K}$, as well as $(\Sigma^z)^{-1}$. Optimal estimation of these quantities can be performed under the conditions presented and discussed below.

Assumption 2 implies that $X_j$ is $\gamma_x$-subgaussian with $\gamma_x = (\gamma_z \sqrt{B_z} + \gamma_w)$, as shown in Lemma 3 in the supplement, and it is well known that in this case,

$$\mathbb{P}\left\{ \max_{1 \leq j < \ell \leq p} |\Sigma_{j\ell} - \Sigma_{j\ell}| \leq \delta \right\} \geq 1 - (p \vee n)^{-c'}$$

for some constant $c' > 0$ and

$$\delta = c \sqrt{\log(p \vee n)/n}$$

with $c = c(\gamma_x) > 0$ sufficiently large (see, for instance, (Bien et al., 2016, Lemma 1)). The leading constant $c$ can be chosen via cross-validation by the criterion on page 25 of Bing et al. (2017). With this form of $\delta$, the following theorem gives the statistical guarantees of the estimates $\hat{K}$ and $I_k$.

**Theorem 2** (Theorem 3 in Bing et al. (2017)). Under Assumptions 1 and 2, assume $\log p = o(n)$. Then, with probability greater than $1 - (p \vee n)^{-c}$, for some constant $c > 0$, we have

1. $\hat{K} = K$;
2. $I_k \subseteq \hat{I}_{\pi(k)} \subseteq I_k \cup J^k_1$, for all $k \in [K]$.

Here $\pi : [K] \rightarrow [K]$ is a permutation and $J^k_1 := \{j \in J : |A_{jk}| \geq 1 - 4\delta/\nu\}$ with constant $\nu$ defined in Assumption 1.

Since we do not impose any separation condition between $A_I$ and $A_J$, we cannot expect to recover $I$ perfectly in the presence of quasi-pure variables defined in the set $J_1 := \bigcup_{k=1}^K J^k_1$. Indeed, when $\log p = o(n)$, for any $j \in J^k_1$ we have $|A_{jk}| \approx 1$ and $A_{jk'} \approx 0$ for any $k' \neq k$, so variables corresponding to $J^k_1$ are very close to the pure variables in $I_k$, and possibly indistinguishable from one another.

Our estimator $\hat{\beta}$ depends on $\hat{I}$, and its performance is likely affected by the cardinality of $J_1$. Fortunately, if $|J_1|$ is small relative to $|I|$, the influence of misclassified $X$-variables with entries in $J_1$ becomes negligible. We introduce

$$\rho := \max_k \frac{|J^k_1|}{|I_k| + |J^k_1|} \in [0, 1)$$

(17)

to quantify the influence of quasi-pure variables on the quality of our estimation. Theorem 4 shows that optimal estimation of $\beta$ is possible in the presence of quasi-pure variables as long as their number is negligible relative to the number of pure variables in the same group, in that the following assumption holds.

**Assumption 3.** $\rho = O(1/\sqrt{mK})$ as $n \rightarrow \infty$, with $m := \min_{k \in [K]} |I_k|$.

This assumption holds if $\max_k |J^k_1| = O(\sqrt{mK})$. For fixed $K$ and $m$, we expect $|J^k_1| \rightarrow 0$ and $|J_1| \rightarrow 0$ whenever $\log p = o(n)$, while Assumption 3 merely requires $|J^k_1| = O(1)$. For $K \rightarrow \infty$, we allow the number of quasi-pure variables $|J_1|$ to grow, but no faster than $O(\sqrt{mK})$.

Another quantity that needs to be controlled is the covariance matrix $\Sigma^z$. It plays the same role as the Gram matrix in classical linear regression with random design.
Assumption 4. Assume $C_{\text{min}} < \lambda_{\text{min}}(\Sigma^z) \leq \lambda_{\text{max}}(\Sigma^z) < C_{\text{max}}$ for some constants $C_{\text{min}}$ and $C_{\text{max}}$ bounded away from 0 and $\infty$.

These assumptions, in addition to Assumption 1, allow for a cleaner presentation of our results. We can trace explicitly the dependency of the estimation rate for $\beta$ on $\rho$, $C_{\text{min}}$ and $C_{\text{max}}$ in the proofs. An important feature of this framework is that under Assumptions 1 – 4 and

$$K = O\left(n/\log(p \lor n)\right),$$

the matrix $\hat{\Sigma}^z$ satisfies the strict inequality

$$\|\hat{\Sigma}^z - \Sigma^z\|_{\text{op}} < \lambda_{\text{min}}(\Sigma^z),$$
in probability, as $n \to \infty$.

This is proved in Lemma 11 of the supplement and in turn guarantees, via Weyl’s inequality, that $\hat{\Sigma}^z$ can be inverted, with probability tending to 1.

2.4 Minimax lower bounds for estimators of $\beta$ in Essential Regression

To benchmark our estimator of $\beta$, we derive the minimax optimal rate of $\|\hat{\beta} - \beta\|$ over the parameter space $(\beta, \Sigma^z, A) \in S(R, m)$ with

$$S(R, m) := \left\{ (\beta, \Sigma^z, A) : \|\beta\| \leq R, \Sigma^z \text{ satisfies Assumption 4}, \right.$$  

$$A \text{ satisfies Assumption 1 with } \min_k |I_k| = m \right\},$$

where $I_k$ is defined in (9). For the purpose of the minimax result, it suffices to consider the joint distribution of $(X, Y)$ as

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_{p+1}\left(0, \begin{bmatrix} A\Sigma^z A^T + \tau^2 I_p & A\Sigma^z \beta \\ \beta^T \Sigma^z A^T & \sigma^2 \end{bmatrix}\right)$$

for $(\beta, \Sigma^z, A) \in S(R, m)$ and some positive constants $\sigma^2$ and $\tau^2$.

Theorem 3. Let $K \leq \bar{c}(R^2 \lor m)n$ for some positive constant $\bar{c}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random variables from the normal distribution in (20). Then, there exist constants $c' > 0$, $c'' \in (0, 1]$ depending only on $\bar{c}$, $C_{\text{max}}$, $C_{\text{min}}$, $\sigma^2$ and $\tau^2$, such that

$$\inf_{\hat{\beta}} \sup_{(\beta, \Sigma^z, A) \in S(R, m)} \mathbb{P}\left\{\|\hat{\beta} - \beta\| \geq c' \left(1 \lor \frac{R}{\sqrt{m}}\right) \cdot \sqrt{\frac{K}{n}}\right\} \geq c''.$$

The inf is taken over all estimators $\hat{\beta}$ of $\beta$.

To the best of our knowledge, this is a new result in the factor regression model literature and it is interesting to place our results in a broader, related, context. For this, note that under the Essential Regression framework, if $I$ and $A_I$ were known, the pure variable assumption implies

$$Y = Z^T \beta + \varepsilon, \quad X_I = Z + W_I$$

with $X_I := \Pi_I^T X := (A_I^T A_I)^{-1} A_I^T X_I$ and $W_I = \Pi_I^T W_I$. The covariance structure of the error term $W_I$ is diagonal, as $\text{Cov}(W_I) = \Pi_I^T \Sigma^w_I \Pi_I$, when $\Sigma^w = \tau^2 I_p$ and $|I_k| = m$ for all $k \in [K]$. 

11
When \( I \), and thus \( A_I \), are known, \( \bar{X}_I \) is observable, and model (22) becomes an instance of an errors in variables model. The minimax optimal lower bound for estimating \( \beta \) in such models has been derived recently in Belloni et al. (2017), under sparsity assumptions on \( \beta \). In the particular case of non-sparse \( \beta \), which we treat here, their lower bound agrees with that given by our Theorem 3, although their bound is derived over a larger class, and can only be compared with (21) when \( I \) is known.

Theorem 3 above shows that, from the point of view of estimating \( \beta \) consistently in Essential Regression and, as we will show below, also for consistent prediction, the number of factors \( K \) can grow with \( n \), a scenario not treated in the more classical factor regression literature summarized in Section 3.

It also reveals that, as in the classical regression set-up, although \( K \) can grow with \( n \) in Essential Regression, consistent estimation of unstructured \( \beta \) cannot be guaranteed when \( K > n \). We discuss briefly estimation of \( \beta \) under sparsity assumptions in Remark 3 of the next section, but a full analysis of this scenario is beyond the scope of the current paper, and will be treated in follow-up work.

### 2.5 Consistency of \( \hat{\beta} \) in \( \ell_2 \)-norm: Rates of convergence and optimality

The following theorem states the convergence rate of \( \min_{P \in \mathcal{H}_K} \| \hat{\beta} - P\beta \| \).

**Theorem 4.** Assume Assumptions 1 – 4 hold. Let \( K \log(p \vee n) = O(n) \). Then, with probability greater than \( 1 - \frac{1}{n} \) for some constant \( c > 0 \), the matrix \( \hat{\Sigma}^z \) is non-singular and the estimator \( \hat{\beta} \) given by (14) satisfies:

\[
\min_{P \in \mathcal{H}_K} \| \hat{\beta} - P\beta \| \lesssim \left( 1 \vee \frac{\| \beta \|}{\sqrt{m}} \right) \sqrt{\frac{K \log(p \vee n)}{n}}. \tag{23}
\]

**Remark 1.** The estimator \( \hat{\beta} \) achieves the minimax rate in Theorem 3 up to a \( \log(p \vee n) \) term. Inspection of the proof, when \( K \to \infty \), tells us that the \( \log(p \vee n) \) terms in the condition \( K \log(p \vee n) = O(n) \) and the upper bound in inequality (23) can be improved to \( K \) correspondingly. This additional log \( K \) is the price to pay for not observing \( Z \). On the other hand, comparing (23) with the convergence rate of the least squares estimator (OLS) when \( Z \) is observable, the additional \( \| \beta \| / \sqrt{m} \) is due to the design error \( W \). This additional error becomes negligible when \( \| \beta \| = O(1) \) and \( K = o(m) \). In the worst case scenario the rate of \( \hat{\beta} \) is slower than the OLS by a factor of \( \sqrt{K} \).

**Remark 2.** Expression (11) suggests that, in addition to our proposed estimator \( \hat{\beta} \), one can also consider the following estimator of \( \beta \), that uses the estimated allocation matrix \( \hat{A} \):

\[
\hat{\beta}_{full} = \frac{1}{n} \left( \hat{\Sigma}^z \right)^{-1} \left[ \hat{A}^T \hat{A} \right]^{-1} \hat{A}^T X^T Y. \tag{24}
\]

However, we recommend \( \hat{\beta} \) over \( \hat{\beta}_{full} \) as it has better theoretical and numerical performance. The construction of (24) involves the estimation of the full matrix \( A \in \mathbb{R}^{p \times K} \), while (14) only requires the estimation of the sub-matrix \( A_I \in \mathbb{R}^{|I| \times K} \), with \( |I| \) possibly much smaller than \( p \). As a result, the consistency of \( \hat{\beta} \) can be obtained at a much faster (minimax optimal) rate than that for \( \hat{\beta}_{full} \). This is verified in the simulation study presented in Section 5. Proposition 19 in 6 shows that the rate of \( \hat{\beta}_{full} \) is slower than that of the proposed \( \hat{\beta} \) by a substantial factor of \( \sqrt{(p / K) \vee \| A_I \|} \).
Remark 3. Assumption 4 on the eigenvalues of $\Sigma^z$ becomes increasingly restrictive if $K$ increases. Instead, as in Bing et al. (2017), one can use the notion of $\ell_q$-sensitivities $\kappa_q(\Sigma^z, s)$, that combines sparsity $s = \|\beta\|_0$ of the vector $\beta$ and restricted eigenvalues of $\hat{\Sigma}^z$. We will address this setting in a different paper. As a short preview, we propose in the case of large $\hat{K}$ to estimate $\beta$ by

$$\hat{\beta}_d = \arg \min_{\beta \in \mathbb{R}^K} \left\{ \|\beta\|_1 : \|\hat{\Sigma}^z \beta - \hat{h}\|_\infty \leq \lambda + \mu \|\beta\|_1 \right\}$$

for some parameters $\lambda$ and $\mu$. Explicit forms of $\lambda$ and $\mu$ are given in 6, where we prove that with probability tending to one, as $(p \lor n) \to \infty$, for $1 \leq q \leq +\infty$,

$$\min_{P \in \mathcal{H}_K} \|\hat{\beta}_d - P \beta\|_q \lesssim \left[ \kappa_q(\Sigma^z, s) \right]^{-1} (1 \lor \|\beta\|_1) \sqrt{\log (p \lor n) / n}$$

and the $\ell_q(\Sigma^z, s)$-sensitivities can be bounded via the inequality

$$\left[ \kappa_q(\Sigma^z, s) \right]^{-1} \leq s^{1/q} \|\Sigma^z\|^{-1}_\infty, 1.$$  

Hence, $\hat{\beta}_d$ adapts to the unknown sparsity $s$ of $\beta$. Finally, using the connection between our model and the errors in variables model mentioned in Section 2.4 above, one can adapt the conic programming approach of Belloni et al. (2017) to construct an alternative sparse estimator, that can be shown to attain the same rate as in (26), but at increased computational cost. A full analysis of sparse estimators in Essential Regression is deferred to future work.

2.6 Inference for $\beta$: Asymptotic normality of $\hat{\beta}$

In this section, for ease of presentation, we assume that the signed permutation matrix $P$ is identity and we consider $\Sigma^w = \tau^2 I_p$ and $|I_k| = m$ for all $k \in [K]$, but our proof holds for the general case.

The component-wise asymptotic normality of $\hat{\beta}$ is proved under the challenging, but realistic, scenario in which some of the non-pure variables are very close to the pure variables, justifying their name, quasi-pure variables, introduced in Section 2.3 above. Allowing for this situation is similar to relaxing the signal strength conditions used in the literature on support recovery. In our context, they would correspond to requiring that the pure and non-pure variables are well separated, in that

$$\min_{j \in J} \min_{P \in \mathcal{H}_K} \|A_{j_1} - Pe_1\|_1 \geq c \sqrt{\log (p \lor n) / n}$$

for some universal constant $c > 0$, an assumption that we do not make. We note that it would be equivalent to requiring $\rho = 0$, for $\rho$ defined in (17).

In Section 2.3 we established the convergence rate of $\hat{\beta}$ when $\rho \neq 0$, but satisfies Assumption 3. For the asymptotic normality result, we can still allow $\rho \neq 0$, but require it to be of the smaller size given in Assumption 3'.

Assumption 3'. $\rho = o(1/\sqrt{mK^2 \log (p \lor n)})$ as $n \to \infty$.

The assumption is implied by $\max_k |J^k_1| = o(\sqrt{mK^2 \log (p \lor n)})$, which in turn implies that the number $|J^1|$ of quasi-pure variables is of order $o(\sqrt{m/ \log (p \lor n)})$.  

**Theorem 5.** Under Assumptions 1, 2, 3, 4 and $K \log(p \lor n) = o(\sqrt{n})$, $\hat{\Sigma}^z$ is non-singular with probability tending to one, and for any $1 \leq k \leq K$,

$$\sqrt{n}Q^{-1/2}_{kk}(\hat{\beta}_k - \beta_k) \overset{d}{\to} N(0, 1), \quad \text{as } n \to \infty,$$

where

$$Q_{kk} = \left(\sigma^2 + \frac{\tau^2}{m} \|\beta\|^2 \right) \left(\Omega_{kk} + \frac{\tau^2}{m} \|\Omega_k\|^2 \right) + \frac{\tau^4}{m^2(m - 1)} \sum_{a=1}^K \beta_a^2 \Omega_{ka}^2.$$

**Remark 4.** The proof of Theorem 5 is based on an application of the Lyapunov central limit theorem and on the appropriate control of two remainder terms, given below, via Lemmas 7 – 12 stated and proved in 6. Specifically, in display (B.27) of the supplement, we prove, for any $k$ and on the appropriate control of two remainder terms, given below, via Lemmas 7 – 12 stated and proved in 6. Specifically, in display (B.27) of the supplement, we prove, for any $k \in [K]$, the decomposition

$$\sqrt{n}(\hat{\beta}_k - \beta_k) = \frac{1}{\sqrt{n}} \sum_{t=1}^n \Omega_{kk}^T U_t + \sqrt{n} \left[ \text{Rem}_1 \right]_k + \sqrt{n} \left[ \text{Rem}_2 \right]_k$$

and show that $(nQ_{kk})^{-1/2} \sum_{t=1}^n \Omega_{kk}^T U_t$ converges, in distribution, to $N(0, 1)$. The first remainder term is associated with errors induced by the presence of quasi-pure variables. It is exactly zero, if no such variables exist or, equivalently, if $\rho = 0$. We do allow for $\rho \neq 0$ and impose instead Assumption 3’. Our analysis reveals that the scaled version $\sqrt{n}/Q_{kk}[\text{Rem}_1]_k$ still vanishes asymptotically. The second remainder term is associated with the error of estimating $(\Sigma^z)^{-1}$ by $(\hat{\Sigma}^z)^{-1}$. Under no sparsity conditions on $(\Sigma^z)^{-1}$, condition $K \log(p \lor n) = o(\sqrt{n})$ ensures $\sqrt{n}/Q_{kk}[\text{Rem}_2]_k = o_p(1)$, as desired.

To benchmark the order of magnitude of our asymptotic variance of $\hat{\beta}_k$, we calculate the Fisher information matrix with respect to $\beta$ for a known choice of $A$, $\Sigma^z$, $\tau^2$ and $\sigma^2$.

**Proposition 6.** Consider (20) with $A$ and $\Sigma^z$ chosen from $\mathcal{S}(R, m)$. Let $G = \Omega + \tau^{-2} A^T A$ and $c_0 := \tau^2/\lambda_K(A \Sigma^z A^T)$. For any unbiased estimator $\hat{\beta}$ and $k \in [K]$,

$$\text{Var}(\beta_k) \geq \frac{1}{1 + 3c_0} (\sigma^2 + \beta^T G^{-1}\beta) \left[\Omega_{kk} + \tau^2 \Omega_{kk}^T (A^T A)^{-1} \Omega_k \right].$$

In particular, when $\lambda_1(A^T A) = c_1 m$ for some $c_1 \geq 1$ and $c_0 = \tau^2/(mc_{\text{min}})$, we have, for any $k \in [K]$,

$$\text{Var}(\beta_k) \geq \frac{1}{1 + 3c_0} \left(\sigma^2 + \frac{\tau^2}{(c_0 + c_1)m} \|\beta\|^2 \right) \left(\Omega_{kk} + \frac{\tau^2}{c_1 m} \|\Omega_k\|^2 \right).$$

When (i) $\lambda_1(A^T A) = o(\lambda_1(A^T A))$ (non-pure rows $A_j$ have small signal); and (ii) $\max_k |I_k| \times \min_k |I_k|$ (the pure variables are balanced), we have $\lambda_1(A^T A) = O(m)$ and the asymptotic variance of our estimator $\hat{\beta}$ agrees with the Cramér-Rao lower bound, in order of magnitude, even though $\tau^2$, $\sigma^2$, $A$ and $\Sigma^z$ are unknown.

In practice, one needs to estimate $Q_{kk}$ in order to construct a valid confidence interval for a coordinate of $\beta$. We use the plug-in estimator $\hat{Q}_{kk}$ by replacing $\sigma^2$, $\tau^2_i := \Sigma_{ii}^w$, $|I_k|$, $\beta$ and $\Omega$ by their estimates. Specifically, we use $\hat{\Omega} = (\hat{\Sigma}^z)^{-1}$ and

$$\tau^2_i = \frac{1}{n} X_i^T X_i - \hat{A}_i^T \hat{\Sigma}^z \hat{A}_i, \quad \text{for all } i \in [p];$$

$$\hat{\sigma}^2 = \frac{1}{n} Y^T Y - 2\hat{\beta}^T \hat{h} + \hat{\beta}^T \hat{\Sigma}^z \hat{\beta}.$$
If $\hat{\tau}^2$ or $\hat{\sigma}^2$ is negative, we set it to 0. Armed with these estimates, the following proposition shows that the plug-in estimator $\hat{Q}_{kk}$ consistently estimates the asymptotic variance $Q_{kk}$ of $\hat{\beta}_k$.

**Proposition 7.** Under Assumptions 1 – 4 and $K \log(p \vee n) = o(n)$, we have

$$\left| \frac{\hat{Q}_{kk}^{1/2}}{Q_{kk}^{1/2}} - 1 \right| = o_p(1).$$

Consequently, if additionally Assumption 3' and $K \log(p \vee n) = o(\sqrt{n})$ hold, we have

$$\sqrt{n}\hat{Q}_{kk}^{-1/2}(\hat{\beta}_k - \beta_k) \xrightarrow{d} N(0, 1), \quad \text{as } n \to \infty, \quad k \in [K].$$

To place our inference results in the context of regression problems with an embedded $K$-dimensional structure, we compare the nature of our results with those obtained in $K$-sparse regression models on $p$ variables. Within those, we distinguish between two inferential goals: (i) Inference in $K$ dimensions and (ii) Inference in $p$ dimensions.

To meet goal (i), a standard approach is to employ a method of selecting $\hat{K}$ of the $p$ variables and provide conditions that guarantee that $\hat{K} = K$ and that the selected set coincides with the true variable subset. On the positive side, under such conditions, inference can be performed in the non-random dimension $K$, see, for instance, Bunea (2004). However, this analysis requires that (a) $\min_j |\beta_j| \gg \sqrt{\log p/n}$, the minimax-optimal threshold for support recovery; and (b) the set of the assumed true $K$ variables can be only weakly correlated, as shown in, for instance, Bunea (2008); Bühlmann and van de Geer (2011); Giraud (2015). Since inferential tools are most needed for near-zero values of the regression coefficients, and typically the active variables are correlated, these restrictions on the parameter space and on the covariates render inference after consistent support recovery, such as uniform (in $\beta_j$) confidence intervals, problematic for practical use.

To meet goal (ii) and alleviate the issues posed by the above approach, one continues to select $\hat{K}$ of the $p$ regression variables and estimate their associated regression coefficients, but follows this by a one-step correction (Bickel, 1975) that leads to an estimator in $p$-dimensions. This estimator can be used for component-wise inference of the $p$ true regression coefficients, valid uniformly over $\mathbb{R}^p$. This approach no longer insists on consistent support recovery and the assumptions on the design matrix can be relaxed considerably. Without further assumptions, this approach is valid when, out of the total $p$ measured $X$ variables, only $K \lesssim \sqrt{n}$ are active in the model. This program has been developed by, among others, Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014); Dezeure et al. (2017); Zhang and Cheng (2017); Javanmard and Montanari (2018).

Inference in Essential Regression reaches a compromise between (i) and (ii).

1. It provides valid inference in $K \lesssim \sqrt{n}$ dimensions, but $K$ is the dimension of the vector of latent factors $Z$, and in E-Regression all $p$ regression variables contribute to the response $Y$, albeit indirectly.

2. Theorem 5 uses the fact that $K$ and the partition of pure variables can be consistently estimated. The latter is proved in Theorem 2, which holds under sufficient conditions that do not involve $\beta$, and therefore is uniform over $\beta$, thus alleviating the issues posed by approach (i). Moreover, the conditions of Theorem 2 allow for highly correlated $X$-covariates, unlike the conditions required by approach (i), and to a lesser extent (ii).

Collectively, 1 and 2 summarize the net gain of performing inference at the interpretable $Z$-resolution level relative to inference at the $X$-level, via sparse regression, whenever the former is of interest.
3 Prediction with Essential Regression

The challenge of predicting $Y = Z^T \beta + \varepsilon$ can be met via a two-step procedure, which at the population level is as follows: First predict $Z$ from the observable $X$ to construct $\tilde{Z}$, then predict $Y$ by $\tilde{Y}$ from $\tilde{Z}$. To predict $Z$, recall that model (2) implies

$$X = Z + W,$$

with $X := \Pi^T X := (A^T A)^{-1} A^T X$ and $W := \Pi^T W$, which suggests the construction of the best linear predictor (BLP) of $Z$ from $X$ given by

$$\tilde{Z} = \text{Cov}(Z, X)\text{Cov}(X)^{-1}X = \Sigma^z (\Sigma^z + \Pi^T \Sigma^w \Pi)^{-1} X.$$ 

We make the simple observation that

$$\beta = \arg \min_{\alpha} E(Y - Z^T \alpha)^2 = \arg \min_{\alpha} E(Y - \tilde{Z}^T \alpha)^2,$$

which justifies predicting $Y$ by $\tilde{Y} = \tilde{Z}^T \beta$. A further exploitation of (10) simplifies $\tilde{Y}$ to

$$\tilde{Y} = X^T \Pi [\Sigma^z + \Pi^T \Sigma^w \Pi]^{-1} \Sigma^z \beta = X^T A \text{Cov}(A^T X)^{-1} \text{Cov}(A^T X, Y).$$

Based on this population level expression of $\tilde{Y}$, we use the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to predict $Y$ by

$$\hat{Y} = X \hat{A} \left[ \hat{A}^T X^T X \hat{A} \right]^{-1} \hat{A}^T X^T Y := X \hat{\theta}$$

where $M^+$ denotes the Moore-Penrose inverse of matrix $M$ and $\hat{A}$ is defined in display (A.2) – (A.4) of Appendix A in the supplement.

The prediction for a single new data point $X_*$ (with corresponding $Z_*$) is given by

$$\hat{Y}_* = X_*^T \hat{\theta} = X_*^T \hat{A} \left[ \hat{A}^T X^T X \hat{A} \right]^{-1} \hat{A}^T X^T Y.$$ (33)

In the following theorem, we present bounds for the in-sample prediction risk $E[\|\hat{Y} - Z^T \beta\|^2]$ and the new-data prediction risk $E[(\hat{Y}_* - Z_*^T \beta)^2 \wedge T]$ for all $T > 0$. We need some additional notation. We write $C_W = \max_{i \in [p]} \Sigma^w_i, \kappa_A = \lambda_1(A^T A) / \lambda_K(A^T A), \Lambda_{\min} = \lambda_K(A^T A)$, and define

$$R_A = \|A_J\|_0 \frac{\log(p \vee n)}{n}$$

(34)

**Theorem 8.** Under Assumptions 1, 2, 4 and $K \log(p \vee n) = O(n)$, we have

$$\frac{1}{n} E \left[ \|\hat{Y} - Z^T \beta\|^2 \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2}{\Lambda_{\min}} C_W \left\{ 1 + \kappa_A R_A + \frac{R_A^2}{\Lambda_{\min}} \right\}.$$ (35)

Additionally, if $R_A / \Lambda_{\min} = o(1)$ holds, we have

$$\frac{1}{n} E \left[ \|\hat{Y} - Z^T \beta\|^2 \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2}{\Lambda_{\min}} C_W (1 + R_A),$$ (36)

and, for any $T > 0$,

$$E \left[ (\hat{Y}_* - Z_*^T \beta)^2 \wedge T \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\sigma^2 \vee \|\beta\|^2}{\Lambda_{\min}} C_W (1 + R_A).$$ (37)
We explain this result below, and defer a detailed discussion on simplifications of the bound to Section 4, where we connect prediction with Essential Regression to prediction in regression with clustered variables.

Remark 5 (Oracle prediction risk). We emphasize that the loss functions in Theorem 8 differ from \( \|Z\beta - Z\hat{\beta}\|^2 \), which corresponds to the unobserved matrix \( Z \), in which case \( Z\hat{\beta} \) is not bona fide predictor of \( Y \). A bound on this loss can be easily derived from Theorem 4:

\[
\frac{1}{n} \|Z\beta - Z\hat{\beta}\|^2 \lesssim \left( 1 + \frac{\|\beta\|^2}{m} \right) \frac{K \log(p \lor n)}{n},
\]  

with probability tending to one. This inequality is proved in Section B.4 of the supplement. Bound (38) is overly optimistic, as it obscures the main difficulty in prediction via factor regression models, that of predicting the unobserved \( Z \). We have included this bound as a benchmark, and for completeness of exposition.

Remark 6 (Interpretation of the prediction rate). The first term \( K\sigma^2/n \) in the right-hand side of (36) is the typical prediction error of \( Y \) in \( K \) dimensions, had we observed \( Z \), while the second term \((C_W\|\beta\|^2/\Lambda_{\min})(1 + R_A)\) is the price to pay for predicting \( Z \) and estimating \( A \). Lemma 21 in the supplement shows that \( C_W\|\beta\|^2/\Lambda_{\min} \) is the irreducible error of (linearly) predicting \( Z \) via \( \hat{X} \), even at the population level:

\[
\min_{\eta} \mathbb{E} \left[ (\beta^T Z - \eta^T \hat{X})^2 \right] = \mathbb{E} \left[ (\beta^T Z - \hat{\beta}^T \tilde{Z})^2 \right] = O \left( C_W\|\beta\|^2/\Lambda_{\min} \right).
\]

Furthermore, predicting \( Z \) necessitates the estimation of \( A \), and Lemma 15 in the supplement shows that \( \|\hat{A} - A\|^2 = O_p(R_A) \). This is the minimax optimal rate for estimating \( A \) in this model, see Bing et al. (2017).

Remark 7 (Conditions of Theorem 8). The quantity \( \Lambda_{\min} \) is the smallest eigenvalue of the matrix \( A^T A = A_J^T A_I + A_J^T A_J \). The matrix \( A_J^T A_I \) is diagonal with the smallest eigenvalue \( m = \min_k |I_k| \). Hence

\[
\Lambda_{\min} \geq \lambda_{\min}(A_J^T A_J) + \lambda_{\min}(A_J^T A_J) \geq m + \lambda_{\min}(A_J^T A_J)
\]  

using the min-max theorem for the first inequality. A limit case of the condition \( R_A/\Lambda_{\min} = o(1) \) corresponds to \( \lambda_{\min}(A_J^T A_J) = o(m) \), in which case we require that the sub-matrix \( A_J \) corresponding to non-pure variables have support that satisfies \( \|A_J\|_0 \log(p \lor n) = o(nm) \). In Section 3.1, we further contrast this condition, and the results of this section, with conditions and results obtained for prediction in principal component regression. As a last point of comparison, conditions on the sparsity of \( A \) simply mean that not all \( p \) variables in the vector \( X \) contribute to explaining a particular \( Z_k \), for each \( k \), which is the main premise of Essential Regression. However, all \( p \) variables are allowed to participate in prediction, indirectly via \( Z \). Our sparsity requirement is quite different than that in prediction via sparse regression. In the latter case, prediction is ultimately based on a subset of \( X \), and a strong sparsity requirement limits the overall number of variables \( X \) used for prediction, thereby potentially increasing the bias.

Remark 8. We use the bounded loss \( (\tilde{Y}_* - Z\beta)^2 \wedge T \) in (37), with arbitrarily large \( T \), for technical convenience, in order to avoid having infinite prediction risk for a new observation, as \( \hat{A}^T X^T X \hat{A} \) may not invertible, on an event with probability converging to zero. Alternatively, we may set our estimate to zero in such cases, see, for instance, Baraud (2002). Such a truncation will not affect the risk. This problem does not arise for the in-sample prediction risk \( \mathbb{E}[\|\tilde{Y} - Z\beta\|^2] \), since \( \tilde{Y} \) utilizes the projection matrix onto \( X \hat{A} \), given in (32), which is uniquely defined.
We close this section by briefly discussing alternative predictors. A natural candidate employs only $X_I$, the vector of pure variables. Using analogous reasoning as above, included for completeness in 6, the population level predictor has the expression

$$\tilde{Y}_I := X_I^T A_I [\text{Cov}(A_I^T X_I)]^{-1} \text{Cov}(A_I^T X_I, Y),$$

and the in-sample prediction risk of its sample counterpart is $K\sigma^2/n + C_W\|\beta\|^2/m$, as stated and proved in Theorem 20 in Appendix C of the supplement. Although this predictor only involves $A_I$, which can be estimated with lower error than $A$, the irreducible prediction error, arising from predicting $\beta^T Z$ via $X_I$, dominates. Indeed, Lemma 21 in the supplement shows that at the population level

$$0 \leq \mathbb{E}[(\beta^T Z - \tilde{Y}_I)^2] - \mathbb{E}[(\beta^T Z - \hat{Y})^2] = O(C_W\|\beta\|^2/m) .$$

Finally, if, in the argument presented at the beginning of this section, we predict $Z$ from $X$ instead of $\bar{X}$, then we arrive at a weighted (by an estimate $\hat{\Omega}^w$ of $[\Sigma^w]^{-1}$) version of the predictor $\hat{Y}$ in (32), given by

$$\hat{Y}_w = X \hat{\Omega}^w A [\hat{A}^T \hat{\Omega}^w X^T X \hat{\Omega}^w A]^+ \hat{A}^T \hat{\Omega}^w Y .$$

The analysis of its prediction risk is almost identical to that of $\hat{Y}$, and the resulting risk bounds are the same, but require the extra assumption that $\min_{1 \leq i \leq p} \Sigma^{w}_{ii} > 0$.

### 3.1 Comparison with principal-component based prediction

A model-free, widely used, predictor of $Y$ in regression problems with high dimensional, highly correlated, design is based on the first $K$ principal components of $X$ (PCR). As mentioned in the Introduction, its population level expression is

$$Y_{PCR}^* = X^T U_K [\text{Cov}(U_K^T X)]^{-1} \text{Cov}(U_K^T X, Y)$$

(42)

where $U_K$ is the $p \times K$ matrix with columns equal to the $K$ eigenvectors of $\text{Cov}(X)$ corresponding to its $K$ largest eigenvalues. It is interesting to note that, when the factor regression model

$$Y = \beta^T Z + \varepsilon, \quad X = BZ + W,$$

holds, with the same $K$, and when $\Sigma^w = \tau^2 I_p$, for some $\tau$, then

$$Y_{PCR}^* = Y_{FR}^*,$$

where

$$Y_{FR}^* = X^T B [\text{Cov}(B^T X)]^{-1} \text{Cov}(B^T X, Y)$$

(44)

is the best linear predictor of $Y$ from $B^T X$ in any factor regression model (43). We could not find this simple result in the literature, and prove it in the first part of Lemma 9 below. This result suggests that prediction via $K$ principal components can be justified within the factor regression framework, with $K$ factors. We strengthen it, in Lemma 9 below, with conditions under which $Y_{PCR}^*$ and $Y_{FR}^*$ are close, for a general noise matrix $\Sigma^w$. It is clear that, although $Y_{FR}^*$ remains unchanged after we replace $B$ in the previous expression by $\tilde{B} = BQ$ for any invertible matrix $Q$ (keeping $X = BZ + W$), we have $Y_{PCR}^* \neq Y_{FR}^*$ in general. However, these two quantities can be close, depending on the size of the signal-to-noise ratio defined as

$$\xi = \lambda_K(B \Sigma^w B^T) / \|\Sigma^w\|_{op} .$$

(45)
Lemma 9. Under model (43), let $Y_{PCR}^*$ and $Y_{FR}^*$ be defined as above.

1. If $\Sigma^w = \tau^2 I_p$, then $Y_{PCR}^* = Y_{FR}^*$ almost surely.

2. Suppose $B\Sigma^zB^T$ has decreasing eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_K$ and

$$\min_{1 \leq k \leq K-1} (\lambda_k - \lambda_{k+1}) \geq c_* \lambda_K, \quad \xi > 1, \quad \xi \to \infty$$

for some constant $c_* > 1$. Then there exists an invertible matrix $Q$ such that

$$\|U_K - BQ\|^2 = O(K/\xi^2).$$

Furthermore, if (46) holds with some constant $c_* > 2$ and

$$\xi > \kappa \sqrt{K}, \quad \xi/(\kappa \sqrt{K}) \to \infty,$$

with $\kappa := \lambda_1(B\Sigma^zB^T)/\lambda_K(B\Sigma^zB^T)$, we have

$$\mathbb{E} [Y_{PCR}^* - Y_{FR}^*]^2 \lesssim \beta^T \Sigma^z \beta \left(\frac{\kappa K}{\xi}\right)^2.$$

Lemma 9 shows that the model-free $Y_{PCR}^*$ can approximate the best linear predictor of $Y$, in any factor regression model, including Essential Regression, at least at the population level, when either $\Sigma^w$ is proportional to the identity matrix, or the signal-to-noise ratio $\xi \to \infty$ sufficiently fast as $p \to \infty$. The latter is implicitly assumed in the existing literature on factor regression models, see Assumption 5 below.

Under the Essential Regression model, the best linear predictor of $Y$, has the same generic form as (44), with $B = A$, but $A$ has the structure postulated in Assumption 1. The model-tailored, sample-based, predictor (32) defined above makes use of this structure, with the net benefit of a reduction in the prediction risk, which is provably small under weaker signal strength assumptions than those required by PCR-type prediction, in unstructured factor regression models. This is verified by our simulations in Section 5, and formalized below.

If $K$ is fixed and, moreover, known, Stock and Watson (2002a,b) consider model (43), under some additional conditions, and prove that $Y_{FR}^*$ and $Y_{PCR}^*$, and their estimates, are close as $n,p \to \infty$, while Bai and Ng (2006) further provides asymptotic prediction intervals for $Y_{PCR}^*$. The latter are constructed relative to an asymptotically normal estimator of, in our notation, $Q\beta$, for a particular, and unknown, invertible matrix $Q$, which differs considerably from our asymptotic results in Theorem 5 in Section 2.6.

A fully data-driven prediction scheme requires the estimation of $K$, and Bai and Ng (2002) provides several AIC/BIC selection type criteria for this purpose. Lam and Yao (2012); Ahn and Horenstein (2013) propose a different approach, based on the ratio of adjacent eigenvalues of the sample covariance matrix. Under suitable conditions, both approaches consistently select $K$ as $n,p \to \infty$. The second criterion tends to have better finite sample performance, and is the one we use in our simulation comparison in Section 5.

All the aforementioned results are proved under conditions that are commonly used in classical factor models (Chamberlain and Rothschild, 1983; Connor and Korajczyk, 1986; Bai and Ng, 2008; Bai, 2003; Fan et al., 2011, 2013). The main assumptions are summarized below:
Assumption 5. Under model (43), assume $K$ is fixed, $\|\beta\|_\infty = O(1)$, $\|\Sigma^w\|_{op} = O(1)$ and all eigenvalues of $p^{-1}B^TB$ and $\Sigma^z$ are of constant order.

We make the important observation that Assumption 5 assumes that $K$ is fixed and implies that the signal-to-noise ratio $\xi$ in (45) satisfies

$$\xi = \frac{\lambda_K(B\Sigma^zB^T)}{\lambda_1(\Sigma^w)} \asymp p,$$

(48)

where we write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Meanwhile, in the context of Essential Regression, using Assumptions 2 and 4 and the fact that $\Sigma^w$ is diagonal, we have instead

$$\xi = \frac{\lambda_K(A\Sigma^zA^T)}{\lambda_1(\Sigma^w)} \asymp \Lambda_{\min} := \lambda_K(A^TA).$$

Under the structural assumptions on model parameters made by Essential Regression:

(1) $K$ is allowed to grow with $n$ and $p$ and can be estimated consistently (Theorem 2), for sub-Gaussian data and $\log(p) = o(n)$;

(2) precise expressions (35), (36) and (37) for the error bounds elucidate how various quantities in the model, influence the quality of prediction. These quantities are (i) the number of factors $K$ via the norm of $\beta$, (ii) the amount of shared information between the components of $X$, via their common factors $Z$, quantified by the size $s_J = \|A_J\|_0$ of the support of the coefficient sub-matrix $A_J$ corresponding to non-pure variables, (iii) the sample size $n$, and (iv) the signal-to-noise ratio $\xi$.

To illustrate the second point, for consistent prediction, (36) in Theorem 8 together with (34) requires that

$$\frac{\xi}{\|\beta\|^2[1 \lor \log(p \lor n)s_J/n]} \to \infty, \text{ as } n \to \infty.$$ (49)

In the worst case scenario of $s_J \asymp Kp$ (no sparsity in $A_J$) and $\|\beta\|^2 = O(K)$ (assuming $\|\beta\|_\infty = O(1)$ as in Assumption 5), (49) boils down to $\xi \gg PK^2\log(p \lor n)/n$. This required rate for $\xi$ becomes less restrictive if $A_J$ and/or $\beta$ is sparse.

Furthermore, consistent prediction can be established under a smaller signal-to-noise ratio $\xi$ than (48) for PCR. To make a fair comparison, we consider the case that $K$ is fixed and $\|\beta\|_\infty = O(1)$ as assumed in Assumption 5 for PCR, and $A_J$ is not sparse. From (49), consistent prediction in Essential Regression is then guaranteed by

$$\xi \gg p\log(p \lor n)/n.$$

This is a weaker assumption on $\xi$ than (48), required by PCR, a fact verified in our simulation studies in Section 5.2 (see Figure 2).

4 Essential Regression as regression with clustered predictors

The decomposition (2) in our model formulation can be used as a model for possibly overlapping clustering. For this, we interpret $A$ as an allocation matrix that assigns the $X$-variables to possibly overlapping groups $G_k$ corresponding to the components of $Z$ via

$$G_k = \{j \in [p] : A_{jk} \neq 0\}.$$
This approach was first proposed in Bing et al. (2017), and their algorithm, called LOVE, was shown to estimate clusters consistently. With this interpretation, the quantities introduced above, 

$$\tilde{X} := \Pi^T X := (A^T A)^{-1} A^T X, \quad \hat{X}_I := \Pi_I^T X_I := (A_I^T A_I)^{-1} A_I^T X_I,$$

can be viewed as weighted cluster averages and averages of pure variables, respectively. As discussed in Section 1.1.3 of the Introduction, Essential Regression provides a framework within which we can analyze when the commonly used cluster averages can be used for downstream analysis, with statistical guarantees. While these quantities are appropriate for prediction purposes, we need the best linear predictors

$$\tilde{Z} = \text{Cov}(Z, \tilde{X}) [\text{Cov}(\tilde{X})]^{-1} \tilde{X} = \Sigma^z (\Sigma^z + \Pi^T \Sigma^w \Pi)^{-1} \tilde{X}.$$  (50)

$$\tilde{Z}_{(I)} = \text{Cov}(Z, \tilde{X}_I) [\text{Cov}(\tilde{X}_I)]^{-1} \tilde{X}_I = \Sigma^z (\Sigma^z + \Pi_I^T \Sigma_I^w \Pi_I)^{-1} \tilde{X}_I.$$  (51)

of $Z$ from $\tilde{X}$ and $\tilde{X}_I$, respectively, for inference of $\beta$. We remark that, at the population level,

$$\beta = \begin{cases} 
\arg\min_\alpha \mathbb{E} [Y - \alpha^T \tilde{Z}]^2 & \neq \arg\min_\alpha \mathbb{E} [Y - \alpha^T \tilde{X}]^2 \\
\arg\min_\alpha \mathbb{E} [Y - \alpha^T \tilde{Z}_{(I)}]^2 & \neq \arg\min_\alpha \mathbb{E} [Y - \alpha^T \tilde{X}_I]^2
\end{cases}$$

suggesting that estimation of $\beta$ should not be based on cluster (weighted) averages. On the other hand,

$$\min_\eta \mathbb{E} \left[ Y - \eta^T \tilde{Z}_{(I)} \right]^2 = \min_\eta \mathbb{E} \left[ Y - \eta^T \tilde{X}_I \right]^2 \geq \min_\eta \mathbb{E} \left[ Y - \eta^T \tilde{Z} \right]^2 = \min_\eta \mathbb{E} \left[ Y - \eta^T \tilde{X} \right]^2.$$  (52)

Thus, while the minimizers of $\mathbb{E} [Y - \eta^T \tilde{X}]^2$ and $\mathbb{E} [Y - \eta^T \tilde{Z}]^2$ differ, the minimal risk corresponding to $\tilde{X}$ and $\tilde{Z}$ are the same. Moreover, (52) informs us that using $\tilde{X}$ (or $\tilde{Z}$) is to be preferred over $\tilde{X}_I$ (or $\tilde{Z}_{(I)}$).

Summarizing, the practical implications of the above discussion are as follows:

1. Regressing $Y$ onto the vector of cluster averages, be they weighted or unweighted, will not result in a consistent estimator for $\beta$;

2. Regressing $Y$ onto the corrected averages will estimate $\beta$. The estimator $\hat{\beta}$ proposed and analyzed in Section 2.2 has the latter interpretation, in the following sense:

**Lemma 10.** Let $\hat{Z}_{(I)}$ be the plug-in estimator of $\tilde{Z}_{(I)}$. Then $\hat{\beta}$ in (14) minimizes $\|Y - \hat{Z}_{(I)} \hat{\beta}\|^2$.

We can view the estimator of $\beta$ as a post-clustering estimator, and the results of Theorem 4 as post-clustering inference, at the $Z$-level.

3. For prediction, one should project onto $\tilde{X}$ (or, equivalently, $\tilde{Z}$) rather than $\tilde{X}_I$ (or $\tilde{Z}_{(I)}$). This is precisely what we propose for $\hat{Y}$ defined in (32) above.

**Remark 9** (Regression with non-overlapping clusters). We distinguish between non-overlapping and overlapping clustering. If the clusters do not overlap, then each component of $X$ is related to only one component of $Z$, so $I = \{ p \} = \{ 1, \ldots, p \}$ and $A = A_I \in \{ 0, 1 \}^{p \times K}$, and therefore $\tilde{X}_I = \tilde{X}$ and $\tilde{Z}_{(I)} = \tilde{Z}$. The clusters are given by the uniquely defined partition of $I$, namely $G_k = I_k$ for $k \in [K]$. This model was first proposed and analyzed in Bunea et al. (2019) for a different algorithm, but
the LOVE algorithm adapts to the non-overlapping setting. If the clusters are balanced in that \( \min_k |I_k| \asymp \max_k |I_k| \), and the regression coefficients are bounded, \( \|\beta\|_{\infty} = O(1) \), the convergence rate of \( \hat{\beta} \) in Theorem 4 becomes

\[
\min_{P \in \mathcal{H}_K} \|\hat{\beta} - P\beta\| \lesssim \left( 1 + \sqrt{\frac{K}{n}} \right) \sqrt{\frac{K \log(p \vee n)}{n}}.
\]

Large values of \( p \) relative to \( K \) help in this case, and if \( K^2 \log(p \vee n) = O(p) \), the above rate is essentially the rate of the least squares estimator obtained by regressing \( Y \) on an observable \( Z \), although in our setup \( Z \) is not observed. The in-sample prediction error of \( \hat{Y} \) can be further simplified in the non-overlapping cluster setting with balanced clusters. In this case, we obtain

\[
\frac{1}{n} E \left[ \|\hat{Y} - Z\beta\|^2 \right] \lesssim \frac{K}{n} + \frac{K^2}{p}.
\]

Again, large \( p \) helps – a phenomenon inherent to prediction in factor models.

**Remark 10** (Regression with overlapping clusters). Different phenomena are exhibited when clusters overlap. By Lemma 10, our proposed estimator \( \tilde{\beta} \) is the least squares estimator of \( Y \) on \( \tilde{Z}^{(I)} \).

Another possibility is \( \tilde{\beta} = \arg\min_{\beta} \|Y - \tilde{Z}\beta\|^2 \), the least squares estimator relative to a plug-in estimator \( \tilde{Z} \) of the population-level \( Z \) given in (50). However, we advise against this estimator, as the performance of \( \tilde{\beta} \) is in general worse than that of \( \hat{\beta} \). A simple explanation is that, similarly to \( \tilde{\beta}_{full} \), such an estimate is based on an estimate of \( A^T A \), which in turn accumulates the noise of estimating the full \( p \times K \) matrix \( A \). In contrast, in the estimation of \( \hat{\beta} \) we only require the estimation of the much simpler submatrix \( A_I \). This is further verified in Section 5.1 of the simulation study.

For prediction, the situation is more nuanced, and the degree of overlap is important. Recall that \( J \) is the index set of all variables that belong to more than one group. Again, for simplicity of presentation, we assume that the clusters are balanced in that \( \min_k |I_k| \asymp \max_k |I_k| \) and \( \|\beta\|_{\infty} = O(1) \). This implies that \( m \asymp |I|/K \), and

\[
\Lambda_{\text{min}} \gtrsim \frac{|I|}{K} + |J| \lambda_{\text{min}} \left( \frac{1}{|J|} A_j^T A_j \right) := \frac{|I|}{K} + |J| \psi_J
\]

by (39). We take a closer look at the upper bound

\[
\frac{1}{n} E \left[ \|\hat{Y} - Z\beta\|^2 \right] \lesssim \frac{K}{n} + \frac{\|\beta\|^2}{\Lambda_{\text{min}}} \left\{ 1 + \|A_J\|_0 \frac{\log(p \vee n)}{n} \right\}
\]

\[
\lesssim \frac{K}{n} + \frac{K^2}{|I| + K |J| \psi_J} \left\{ 1 + |J| \max_j \|A_{J_j}\|_0 \frac{\log(p \vee n)}{n} \right\}
\]

for the in-sample prediction risk in Theorem 8. In the moderate overlap scenario of \( |J| = O(|I|) \), or equivalently, \( |I| \asymp p \), and the super-overlap scenario of \( |I| = o(|J|) \), or \( |I| = o(p) \) and \( |J| \asymp p \), this bound reads as

\[
\frac{1}{n} E \left[ \|\hat{Y} - Z\beta\|^2 \right] \lesssim \frac{K}{n} + \frac{K^2}{|I|} + \left( \frac{K^2 \wedge}{\psi_J} \right) \left( \max_j \|A_{J_j}\|_0 \right) \frac{\log(p \vee n)}{n}, \quad \text{moderate-overlap};
\]

\[
\lesssim \frac{K}{n} + \frac{K}{p \psi_J} + \frac{K}{\psi_J} \left( \max_j \|A_{J_j}\|_0 \right) \frac{\log(p \vee n)}{n}, \quad \text{super-overlap}.
\]
Interestingly, large values of $p$ continue to be a blessing in the moderate overlap scenario, but they may become a curse in the super overlap scenario, depending on the behavior of $\psi_J$ as $p \to \infty$.

5 Simulations

In this section, we complement and support our theoretical findings with simulations, focusing on two main areas:

1. The $\ell_2$ convergence rate of $\hat{\beta}$ and the coverage of its 95% confidence interval (CI) of $\beta$;
2. The prediction performance of candidate predictors.

Data generating mechanism: We first describe how we generate $A, \Sigma^z, \Sigma^w$, and $\beta$. Recall that $A$ can be partitioned into $A_I$ and $A_J$. To generate $A_I$, we set $|I_k| = m$ for each $k \in [K]$ and choose $A_I = I_K \otimes 1_m$, where $\otimes$ denotes the kronecker product. Each row of $A_J$ is generated by first randomly selecting its support with cardinality drawn from $\{2, 3, \ldots, \lfloor K/2 \rfloor\}$ and then by sampling its non-zero entries from $\text{Unif}(0, 1)$ with random signs. In the end, we rescale $A_J$ such that the $\ell_1$ norm of each row is no greater than 1. To generate $\Sigma^z$, we set $\text{diag}(\Sigma^z)$ to a $K$-length sequence from 2.5 to 3 with equal increments. The off-diagonal elements of $\Sigma^z$ are then chosen as $\Sigma^z_{ij} = (-1)^{(i+j)(\Sigma^z_{ii} \wedge \Sigma^z_{jj})}(0.3)^{|i-j|}$ for any $i \neq j \in [K]$. Finally, $\Sigma^w$ is chosen by randomly sampling its diagonal elements from $\text{Unif}(2, 3)$.

Then we generate the $n \times K$ matrix $Z$ and the $n \times p$ noise matrix $W$ whose rows are i.i.d. from $N_K(0, \Sigma^z)$ and $N_p(0, \Sigma^w)$, respectively. Then we set $X = ZA^T + W$ and $Y = Z\beta + \varepsilon$ where the $n$ components of $\varepsilon$ are i.i.d. $N(0, 1)$. For each setting, we repeat generating $(X, Y)$ 100 times and record the corresponding results.

5.1 Convergence rate of $\beta$ and the coverage of the 95% confidence interval

We consider the following two settings:

1. Fix $p = 500$, $K = 10$, $m = 10$, and vary $n \in \{200, 400, 600, 800, 1000\}$;
2. Fix $n = 300$, $p = 500$, $K = 10$ and vary $m \in \{5, 10, 20, 30, 40\}$.

The entries of $\beta$ are independently sampled from $\text{Unif}(1, 2)$. For each setting, we calculate the averaged $\ell_2$ errors $\|\hat{\beta} - \beta\|$ of the following four estimators and report them in Table 1:

- $\hat{\beta}$ constructed in (14);
- $\hat{\beta}_{\text{full}}$ defined in (24);
- $\tilde{\beta} := \arg \min_\beta \|Y - \hat{Z}\beta\|^2$ where $\hat{Z}$ is the plug-in type estimator of (50) by using $\hat{\Sigma}^z$, $\hat{\Sigma}^w$ and $\hat{A}$ obtained from (A.1), (27), (A.2), (A.3) and (A.4) in 6.
- $\hat{\beta}_{\text{oracle}} = (Z^T Z)^{-1} Z^T Y$, the oracle least square estimator by using the true $Z$.

For the inference of $\beta$, we calculate the 95% confidence interval (CI) of $\beta_1$ based on Theorem 5 and Proposition 7. Their coverage is shown in Table 1.

Summary: Naturally, $\hat{\beta}_{\text{oracle}}$ is the best since it uses the true $Z$. Among the other estimators, $\hat{\beta}$ has the closest performance to $\hat{\beta}_{\text{oracle}}$ and outperforms the other two estimators in all cases.
The gap between $\hat{\beta}$ and $\hat{\beta}_{\text{oracle}}$ decreases as $m$ increases. The estimation errors of $\hat{\beta}$, $\tilde{\beta}$ and $\hat{\beta}_{\text{full}}$ decrease as $n$ or $m$ increases. These findings support Theorem 4. The average coverage, over 100 repetitions, is 95% or higher for the 95% confidence intervals, which further supports the results of Section 2.6.

Table 1: $\ell_2$ error of different estimators and the coverage of the 95% CI of $\beta_1$.

|       | $\hat{\beta}$ | $\tilde{\beta}$ | $\hat{\beta}_{\text{full}}$ | $\hat{\beta}_{\text{oracle}}$ | coverage(%) |
|-------|----------------|------------------|-----------------------------|-------------------------------|-------------|
| Vary $n$ with $p = 500$, $m = K = 10$ |               |                  |                             |                               |             |
| $n = 200$ | 0.52       | 0.76             | 1.02                        | 0.16                          | 98          |
| $n = 400$ | 0.32       | 0.51             | 0.75                        | 0.11                          | 99          |
| $n = 600$ | 0.26       | 0.45             | 0.63                        | 0.08                          | 100         |
| $n = 800$ | 0.22       | 0.41             | 0.51                        | 0.08                          | 100         |
| $n = 1000$ | 0.20      | 0.36             | 0.49                        | 0.07                          | 98          |
| Vary $m$ with $n = 300$, $p = 500$, $K = 10$ |       |                  |                             |                               |             |
| $m = 5$    | 0.57       | 0.71             | 1.10                        | 0.12                          | 96          |
| $m = 10$   | 0.35       | 0.56             | 0.83                        | 0.12                          | 99          |
| $m = 20$   | 0.25       | 0.37             | 0.43                        | 0.12                          | 100         |
| $m = 30$   | 0.21       | 0.27             | 0.29                        | 0.12                          | 98          |
| $m = 40$   | 0.20       | 0.22             | 0.22                        | 0.12                          | 98          |

5.2 Prediction with E-Regression: a comparative study

In this section we study the numerical performance of our prediction strategy relative to principal component regression (PCR) and the Lasso. We consider the following candidates:

- ER: the Essential Regression predictor $\hat{Y}_* = X^T\hat{\theta}$ from (32). We also considered the weighted predictor defined in (41) and since these two have similar performance, we only present the result of $\hat{Y}_*$.
- PCR-oracle: the principal component regression (PCR) in (42) using the true $K$;
- PCR: the above PCR with $K$ selected via the criterion proposed in Lam and Yao (2012); Ahn and Horenstein (2013). 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.
- Lasso: the “least absolute shrinkage and selection operator” (Tibshirani, 1996) with the tuning parameter chosen via cross-validation.

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

$$X^T\bar{\theta} = X^T[\text{Cov}(X)]^{-1}\text{Cov}(X,Y) = X^T\Omega^wA[\Omega + A^T\Omega^wA]^{-1}\beta,$$
where \( \Omega = [\Sigma^2]^{-1}, \Omega^w = [\Sigma^w]^{-1} \) and the last equality uses Fact 1 in Appendix B of the supplement. Although \( \hat{\theta} \) is not sparse in general, we observe that \( \|\hat{\theta}\|^2 \leq \beta^T[\Omega + A^T\Omega^wA]^{-1}\beta \). Hence its \( \ell_2 \)-norm may be small if \( \|\beta\| \) is small and/or \( \Lambda_{\min} \) is large. Our simulation design allows for these possibilities.

In Section 5.2.1 below we investigate how the prediction errors of these four procedures change as we vary \( p, K \) and the signal-to-noise ratio (SNR) one at a time.

In Section 5.2.2, we consider a limit case of our model (1), corresponding to \( W \approx 0 \), as this set-up favors PCR-type prediction, and also allows for a better understanding of the behavior of the Lasso-based prediction, under our model specification.

The performance metric is based on the new data prediction risk. To calculate it, we independently generate a new dataset \((X_{\text{new}}, Y_{\text{new}})\) containing \( n \) i.i.d. samples drawn according to our data generating mechanism. The prediction risk of the predictor \( \hat{Y}_{\text{new}} \) is calculated as \( \|\hat{Y}_{\text{new}} - Z_{\text{new}}\beta\|^2 / n \).

### 5.2.1 Varying \( p, K \) and SNR one at a time

In order to vary the signal-to-noise ratio, we use a slightly different generating mechanism for \( A_J \). For each row \( A_j \) of \( A_J \), after randomly selecting its support, with cardinality drawn from \( s_j \in \{2, 3, \ldots, K\} \), we sample its non-zero entries from \( N_{s_j}(0, D) \) with \( \text{diag}(D) = (1/s_j, \ldots, 1/s_j) \) and \( D_{ab} = \frac{\zeta|i-j|}{s_j} \) for any \( a \neq b \) and given parameter \( \zeta \in [0, 1] \). In addition, we sample the diagonal entries of \( \Sigma^w \) from \( \text{Unif}(3, 5) \) and the entries of \( \beta \) from \( \text{Unif}(0, 1) \).

To vary \( p \) and \( K \) one at a time, we first set \( n = 300, K = 10, m = 5 \) and choose \( p \) from \( \{200, 400, 600, 800, 1000\} \), then fix \( n = 300, p = 600, m = 5 \) and vary \( K \) in \{10, 20, 30, 40, 50\}. Both settings use \( \zeta = 0.5 \). The prediction risks of the four predictors listed above are shown in Figure 2.

To vary the signal-to-noise ratio \( \xi = \lambda_K(A\Sigma^2A^T)/\lambda_1(\Sigma^w) \), we fix \( \Sigma^2 \) and \( \Sigma^w \), and generate \( A_J \) by choosing \( \zeta \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99\} \). We set \( n = 300, p = 400, K = 10 \) and \( m = 3 \). For each \( \zeta \), we calculate the SNR and plot the prediction risks of each predictor in Figure 2.

![Figure 2: Prediction risks of different predictors as \( p, K \) and SNR vary separately](image)

**Summary:** When \( W \neq 0 \), we do not expect the Lasso to perform well. It is thus surprising that in some regimes its prediction performance is not worse than the methods tailored to this model. The reason is that \( \|\hat{\theta}\| \) is small in some settings. However, the gap between the Lasso and the other methods increases with the size of \( \|\beta\| \), as further illustrated in Section 5.2.2.
Overall, the prediction error for all four methods deteriorates as \( K \) increases or the SNR decreases. This indicates that prediction becomes more difficult for large \( K \) and small SNR. On the other hand, all methods perform better as \( p \) increases. 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 \( Z \) more accurately. This can be seen from our prediction risk in Theorem 8 by noting that \( \Lambda_{\text{min}} \) increases as \( p \) increases. This phenomenon has been observed in the classical factor (regression) model, see, for instance, Stock and Watson (2002a); Bai (2003); Bai and Ng (2008, 2006); Fan et al. (2013).

Among the four candidates, ER has the smallest prediction error in all settings. Furthermore, PCR fails to detect \( K \) and tends to select \( \hat{K} < K \). It is clear that using \( \hat{K} < K \) leads to a large loss in prediction accuracy. This also indicates that, for principal component regression approaches, detecting \( K \) requires larger SNR than making consistent prediction with true \( K \) given. In the first plot, we are in a moderate SNR regime and as expected from Lemma 9, the PCR-oracle approaches have comparable performance to ER. In the second plot, as \( K \) increases, the advantage of ER becomes considerable, which supports the fact that PCR only has guarantees for fixed \( K \).

Finally, in the third plot, all PCR approaches, including the PCR-oracle, are affected the most by a smaller SNR.

5.2.2 Prediction when \( W \approx 0 \)

In this section, we focus on a particular setting in which \( W \approx 0 \). Note that, when \( W = 0 \), model (1) implies \( X = AZ \) and

\[
Y = X^T A (A^T A)^{-1} \beta + \varepsilon := X^T \eta + \varepsilon \implies \mathbb{E}[Y|X] = X^T \eta.
\]

Although \( \eta \) is not entry-wise sparse in general, we see that its \( \ell_2 \) norm is small when \( \Lambda_{\text{min}} \) is large from the inequality \( \|\eta\|^2 \leq \|\beta\|^2 / \Lambda_{\text{min}} \). We thus expect that taking \( W = 0 \) will improve the performance of the Lasso.

In the following, we first consider \( W = 0 \) and compare different predictors for various choices of \( \eta \). Then, we consider \( W \approx 0 \) by adding some small noise to \( W \) and investigate how the predictors change correspondingly.

**Case 1: \( W = 0 \)** We consider both a low-dimension setting \((n = 600, p = 400)\) and a high-dimensional setting \((n = 200, p = 400)\). For both cases, we set \( K = 10, m = 5 \) and consider an initial \( \beta_0 \) whose entries are drawn from Unif(1,2). We generate \( A, \Sigma^x \) and \( \Sigma^w \) according to the data generating mechanism. Then for a fixed \( A \), we choose \( \eta = A (A^T A)^{-1} (\beta_0 + \Delta) \) with \( \Delta \) in \{1, 3, 5, 7, 9\} (the addition is element-wise). For each \( \Delta \), we calculate the prediction risks of different predictors as shown in Table 2. Since PCR-oracle, PCR and ER have the same prediction error, we only present one of them. In fact, since \( X = AZ \), both PCR and PCR-oracle are regressing \( Y \) on the true \( Z \). So is the ER when \( I \) is estimated consistently.

**Summary:** Since ER regresses \( Y \) on the true \( Z \), as we consistently estimate \( I \), ER works very well regardless of the choice of \( \eta \). On the other hand, the Lasso clearly has inferior performance and the gap becomes considerable as \( \|\eta\| \) increases. This is due to the collinearity among \( X \). Indeed, from \( X = AZ \), we observe that all the pure variables in the same group are perfectly correlated. It is well known that the Lasso isn’t stable in the presence of multi-collinearity. This phenomenon is slightly alleviated in the low-dimension case.
Table 2: Table of prediction risks for different \( \Delta \)

| \( \Delta \) | \( p < n \) | \( p > n \) | \( \| \eta \| \) |
|------------|-----------|-----------|---------|
| \( \Delta = 1 \) | 0.015 0.022 | 0.054 0.076 | 2.1     |
| \( \Delta = 3 \) | 0.016 0.033 | 0.055 0.317 | 3.7     |
| \( \Delta = 5 \) | 0.017 0.251 | 0.054 1.394 | 5.3     |
| \( \Delta = 7 \) | 0.017 1.038 | 0.052 2.651 | 6.9     |
| \( \Delta = 9 \) | 0.016 1.997 | 0.058 4.885 | 8.6     |

**Case 2:** \( W \approx 0 \)  When \( W = 0 \), the collinearity of \( X \) may lead to inconsistent estimation via the Lasso, which becomes highly apparent for large values of \( \| \eta \| \). We investigate whether adding noise to \( W \) changes this phenomenon. We consider the high-dimensional setting, \((n = 200, p = 400)\), set \( \Delta = 5 \) and sample \( W \) from \( N(0, \Sigma_w^\rho) \) for a diagonal matrix \( \Sigma_w^\rho \) with each entry sampled from \( \text{Unif}(0, \rho) \) and \( \rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\} \). The case \( \rho = 0 \) corresponds to \( W = 0 \). Note that the data is still generated from model (1). For each \( \rho \), the prediction risks of different predictors are shown in Figure 3. Adding noise to \( W \) clearly makes all predictors deteriorate. Lasso is outperformed over the entire range of \( \rho \).

![Figure 3: The prediction risks for different values of \( \rho \)](image)

6 Analysis of SIV-vaccine data

We tested E-Regression on a high-dimensional dataset of vaccine-induced humoral immune responses, from a recently published study that demonstrated multiple antibody-centric mechanisms of vaccine-induced protection against SIV (Ackerman et al., 2018), the non-human primate equivalent of HIV. The dataset comprised \( p = 191 \) antibody functional and biophysical properties, including Fc effector functions, glycosylation profiles and binding to Fc receptors. The properties were measured for \( n = 60 \) non-human primates (NHPs). For each NHP, the level of protection offered by the vaccine (number of intra-rectal SIV challenges after which the NHP got infected or
whether the NHP remained uninfected after the maximum number of challenges for the study (12), normalized by the total number of challenges) was used as the outcome \( Y \) we regressed to.

Figure 4: Left panel: The prediction mean square errors of E-Regression (ER), Lasso and Principal Component Regression (PCR) from 50 repetitions of 5-fold cross-validation. Right panel: Two representative clusters with their pure variables. (The overlapping variables between these two clusters are more than the plot shows.)

E-Regression was very accurate at predicting outcome and performed better than Lasso and Principal Component Regression in a 5-fold cross-validation framework as shown in Figure 4. Perhaps more interestingly, E-Regression used cluster centers to predict outcome rather than using a sparse set of predictive variables (for instance, via the Lasso). This framework is especially appropriate as vaccine-induced protection is mediated by coordinated immunological mechanisms, rather than a sparse set of antibody properties (variables). While the sparsity assumption is commonly made to avoid overfitting on a high-dimensional dataset, it merely allows the identification of predictive markers, and not actual mechanisms. To gain mechanistic insights, the authors in the original manuscript Ackerman et al. (2018) reconstructed correlation blocks around the Lasso-selected variables. E-Regression already uses cluster centers instead of individual variables to predict outcome. By identifying cluster centers with significant non-zero regression coefficients, our method directly allows the identification of concerted antibody-centric mechanisms that determine outcome. Further, our clustering algorithm uses the notion of pure and mixed variables, which agrees with the notion that some antibody properties work in tandem with several other properties (mixed variables), while others are part of individual immunological signatures (pure variables) (Bournazos and Ravetch, 2017; Nimmerjahn and Ravetch, 2007). For this dataset, of the 191 variables, our clustering algorithm gave an initial estimator \( \hat{K} = 10 \) of the number of factors. We then used the asymptotic normality of \( \hat{\beta} \), established in Section 2.6, to retain the most relevant factors, given in the above picture, under FDR control. As further practical validation, our algorithm identified the only 3 markers of degranulation of natural killer (NK) cells – CD107a, MIP1b and IFNg as pure variables, and placed them in the same cluster, consistent with a-priori expectation. Most of the glycan features were also assigned to the same cluster. This is illustrated in Figure 4. As this pilot data analysis yielded highly interpretable results, we use it as motivation for a full investigation of
this data set, that will be conducted in subsequent work. For now we conclude that E-Regression can be a very useful tool for regression problems in which the influence on $Y$ of representatives of groups of the $X$-variables is of primary interest, and not directly that of the components of $X$.

Acknowledgements

We are grateful to Jishnu Das for help with the interpretation of our data analysis results. Bunea and Wegkamp are supported in part by NSF grant DMS-1712709. Bing is supported in part by NSF grant DMS-1407600.

Supplementary Material

The supplementary document includes the LOVE algorithm in Bing et al. (2017), all the proofs and auxiliary results.

References

Ackerman, M. E., Das, J., Pittala, S., Broge, T., Linde, C., Suscovich, T. J., Brown, E. P., Bradley, T., Natraj, H., Lin, S., Sass, J. K., O’Keefe, S., Mehta, N., Goodman, D., Sips, M., Weiner, J. A., Tomaras, G. D., Haynes, B. F., Lauffenburger, D. A., Bailey-Kellogg, C., Roederer, M. and Alter, G. (2018). Route of immunization defines multiple mechanisms of vaccine-mediated protection against siv. *Nature Medicine* **24** 1590–1598.

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

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*. University of California Press, Berkeley, Calif.

Arora, S., Ge, R., Halpern, Y., Mimno, D. M., Moitra, A., Sontag, D., Wu, Y. and Zhu, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*.

Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012, IEEE 53rd Annual Symposium*. IEEE.

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

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics* **146** 304 – 317. Honoring the research contributions of Charles R. Nelson.

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

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM: Probability and Statistics* **6** 127146.
Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 939–956.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *Journal of the American Statistical Association* **70** 428–434.

Bien, J., Bunea, F. and Xiao, L. (2016). Convex banding of the covariance matrix. *Journal of the American Statistical Association* **111** 834–845.

Bing, X., Bunea, F., Yang, N. and Wegkamp, M. (2017). Adaptive estimation in structured factor models with applications to overlapping clustering. *arXiv:1704.06977v3*.

Bittorf, V., Recht, B., Re, C. and Tropp, J. A. (2012). Factoring nonnegative matrices with linear programs. *arXiv:1206.1270*.

Blei, D. M. and McAuliffe, J. D. (2007). Supervised topic models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*. NIPS’07, Curran Associates Inc., USA.

Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics* **132** 169 – 194. Common Features.

Bournazos, S. and Ravetch, J. V. (2017). Fc receptor function and the design of vaccination strategies. *Immunity* **47** 224–233.

Bühlmann, P., Rütimann, P., van de Geer, S. and Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference* **143** 1835 – 1858.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer-Verlag Berlin Heidelberg.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32** 898–927.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via $\ell_1$ and $\ell_1 + \ell_2$ penalization. *Electron. J. Statist.* **2** 1153–1194.

Bunea, F., Giraud, C., Luo, X., Royer, M. and Verzele, N. (2019). Model Assisted Variable Clustering: Minimax-optimal Recovery and Algorithms. *Annals of Statistics* to appear.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304.

Connor, G. and Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* **15** 373–394.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26.

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

Dezeure, R., Bühlmann, P. and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719.

Fan, J., Liao, Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356.

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

Fan, J., Xue, L. and Yao, J. (2017). Sufficient forecasting using factor models. *Journal of
Econometrics 201 292 – 306.

GIRAUD, C. (2015). Introduction to High-Dimensional Statistics. Monographs on Statistics and Applied Probability.

HAHN, P. R., CARVALHO, C. M. and MUKHERJEE, S. (2013). Partial factor modeling: Predictor-dependent shrinkage for linear regression. Journal of the American Statistical Association 108 999–1008.

HOTELLING, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. British Journal of Statistical Psychology 10 69–79.

IZENMAN, A. J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Series: Springer Texts in Statistics.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15 2869–2909.

JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the lasso: Optimal sample size for gaussian designs. Ann. Statist. 46 2593–2622.

KELLY, B. and PRUITT, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. Journal of Econometrics 186 294 – 316. High Dimensional Problems in Econometrics.

KENDALL, M. G. (1957). A course in multivariate analysis. Hafner Pub. Co.

KLOPP, O., LU, Y., TSYBAKOV, A. B. and ZHOU, H. H. (2017). Structured Matrix Estimation and Completion. ArXiv e-prints.

LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. Ann. Statist. 40 694–726.

NIMMERJAHN, F. and RAVETCH, J. V. (2007). Fc receptors as regulators of immunity. vol. 96 of Advances in Immunology. Academic Press, 179 – 204.

STOCK, J. H. and WATSON, M. W. (2002a). Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97 1167–1179.

STOCK, J. H. and WATSON, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20 147–162.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58 267–288.

TSYBAKOV, A. B. (2009). Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 1166–1202.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press, 210 – 268.

YU, Y., WANG, T. and SAMWORTH, R. J. (2014). A useful variant of the DavisKahan theorem for statisticians. Biometrika 102 315–323.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76 217–242.

ZHANG, X. and CHENG, G. (2017). Simultaneous inference for high-dimensional linear models. Journal of the American Statistical Association 112 757–768.
A The LOVE algorithm

For the reader’s convenience, we give the specifics of estimating \( I \) and \( K \) developed by Bing et al. (2017).

Algorithm 1 Estimate the partition of the pure variables \( I \) by \( \hat{I} \)

\[
\begin{align*}
1: & \quad \text{procedure PureVar}(\hat{\Sigma}, \delta) \\
2: & \quad \hat{I} \leftarrow \emptyset. \\
3: & \quad \text{for all } i \in [p] \text{ do} \\
4: & \quad \hat{I}^{(i)} \leftarrow \{ l \in [p] \setminus \{i\} : \max_{j \in [p] \setminus \{i\}} |\hat{\Sigma}_{ij}| \leq |\hat{\Sigma}_{il}| + 2\delta \} \\
5: & \quad Pure(i) \leftarrow \text{True}. \\
6: & \quad \text{for all } j \in \hat{I}^{(i)} \text{ do} \\
7: & \quad \text{if } |\hat{\Sigma}_{ij}| - \max_{k \in [p] \setminus \{j\}} |\hat{\Sigma}_{jk}| > 2\delta \text{ then} \\
8: & \quad Pure(i) \leftarrow \text{False}, \\
9: & \quad \text{break} \\
10: & \quad \text{if } Pure(i) \text{ then} \\
11: & \quad \hat{I}^{(i)} \leftarrow \hat{I}^{(i)} \cup \{i\} \\
12: & \quad \hat{I} \leftarrow \text{Merge}(\hat{I}^{(i)}, \hat{I}) \\
13: & \quad \text{return } \hat{I} \text{ and } \hat{K} \text{ as the number of sets in } \hat{I} \\
14: & \quad \text{function Merge}(\hat{I}^{(i)}, \hat{I}) \\
15: & \quad \text{for all } G \in \hat{I} \text{ do} \quad \triangleright \hat{I} \text{ is a collection of sets} \\
16: & \quad \text{if } G \cap \hat{I}^{(i)} \neq \emptyset \text{ then} \quad \triangleright \text{Replace } G \in \hat{I} \text{ by } G \cap \hat{I}^{(i)} \\
17: & \quad G \leftarrow G \cap \hat{I}^{(i)} \\
18: & \quad \text{return } \hat{I} \\
19: & \quad \hat{I}^{(i)} \in \hat{I} \quad \triangleright \text{add } \hat{I}^{(i)} \text{ in } \hat{I} \\
20: & \quad \text{return } \hat{I}
\end{align*}
\]

Next, for each \( a \in [\hat{K}] \) and \( b \in [\hat{K}] \setminus \{a\} \), we compute

\[
\hat{\Sigma}_a^z = \frac{1}{|\hat{I}_a|(|\hat{I}_a| - 1)} \sum_{i,j \in \hat{I}_a, i \neq j} |\hat{\Sigma}_{ij}|, \quad \hat{\Sigma}_{ab}^z = \frac{1}{|\hat{I}_a||\hat{I}_b|} \sum_{i \in \hat{I}_a, j \in \hat{I}_b} \hat{A}_{ia}\hat{A}_{ib}\hat{\Sigma}_{ij},
\]

(53)
to form the estimator \( \hat{\Sigma}^z \) of \( \Sigma^z \).

Furthermore, we restate the estimation of \( A_I \) in Bing et al. (2017). For each \( k \in [\hat{K}] \) and the estimated pure variable set \( \hat{I}_k \),

Pick an element \( i \in \hat{I}_k \) at random, and set \( \hat{A}_i = e_k \);

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

(54)
(55)

For the estimation of \( A_J \), we use the Dantzig-type estimator \( \hat{A}_D \) proposed in Bing et al. (2017) given by

\[
\hat{A}_j = \arg \min_{\beta_j} \left\{ \| \beta_j \|_1 : \| \hat{\Sigma}^z \beta_j - (\hat{A}_I^T \hat{A}_I)^{-1} \hat{A}_I^T \hat{\Sigma}_{Ij} \|_\infty \leq \mu \right\}
\]

(56)
for any \( j \in \mathcal{I} \), with tuning parameter \( \mu = O(\sqrt{\log(p \lor n)/n}) \). The estimator \( \hat{A} \) enjoys the optimal convergence rate of \( \max_{j \in [p]} \| \hat{A}_j - A_j \|_q \) for any \( 1 \leq q \leq \infty \) (Bing et al., 2017, Theorem 5).

### B Main proofs

#### B.1 Proof of Proposition 1: the identifiability of \( \beta \)

From the structure of \( \Sigma \) together with Assumption 1, Theorem 1 in Bing et al. (2017) can be directly invoked to show that \( I \) and its partition \( \mathcal{I} \) are identifiable up to a label permutation. In addition, \( A \) is identifiable up to a signed permutation. Suppose we identify \( \tilde{A} = A P \) for some signed permutation \( P \) and, in particular, \( \tilde{A}_I = A_I P \).

First observe that for any \( a, b \in [K] \), \( \Sigma^{z}_{ab} \) is recovered by

\[
\Sigma^{z}_{ab} = A_{ia} A_{jb} \Sigma_{ij}, \quad \text{for } i \in \mathcal{I}_a, \ j \in \mathcal{I}_b, \ i \neq j. \tag{57}
\]

We prove this as follows. Since \( \Sigma = A \Sigma^{z} A^T + \Sigma^w \) with \( \Sigma^w \) diagonal,

\[
\Sigma_{ij} = \sum_{c,d=1}^{K} A_{ia} \Sigma^{z}_{cd} A_{jc} = A_{ia} A_{jb} \Sigma^{z}_{ab},
\]

where in the second step we use that \( i \in \mathcal{I}_a, \ j \in \mathcal{I}_b \). Then (57) follows from \( |A_{ia}| = |A_{jb}| = 1 \).

From \( \tilde{A} \) and the corresponding partition \( \{ \mathcal{I}_a \}_{a \in [K]} \), we can define \( \tilde{\Sigma}^{z} \) to be the matrix with elements \( \tilde{\Sigma}^{z}_{ab} = \tilde{A}_{ia} \tilde{A}_{jb} \Sigma_{ia,jb} \) for some \( i_a \in \tilde{\mathcal{I}}_a, \ j_b \in \tilde{\mathcal{I}}_b \), and \( i_a \neq j_b \). Let \( \pi : [K] \to [K] \) be the permutation mapping corresponding to \( P \). Specifically, \( \pi(a) \) equals the unique \( b \in [K] \) such that \( |P_{ba}| = 1 \). Then for any \( i \in [p] \) and \( a \in [K] \), \( A = AP \) implies \( \tilde{A}_{ia} = P_{\pi(a)} A_{\pi(a)} \) and \( \tilde{I}_a = I_{\pi(a)} \), so

\[
\tilde{\Sigma}^{z}_{ab} = P_{\pi(a)} A_{\pi(b)} P_{\pi(b)} \Sigma_{ia,jb} = P_{\pi(a)} A_{\pi(b)} \Sigma_{ia,jb} = P_{\pi(a)} A_{\pi(b)},
\]

where we use (57) in the second equality since \( i_a \in I_{\pi(a)}, \ j_b \in I_{\pi(b)}, \ i_a \neq j_b \). This shows \( \tilde{\Sigma}^{z} = P^T \Sigma^{z} P \). Finally, we have

\[
\tilde{\beta} = (\tilde{\Sigma}^{z})^{-1} (\tilde{A}^T_{I} \tilde{A}_{I})^{-1} \tilde{A}^T_{I} \text{Cov}(X_I, Y) = (\tilde{\Sigma}^{z})^{-1} (\tilde{A}^T_{I} \tilde{A}_{I})^{-1} \tilde{A}^T_{I} A_I \Sigma^{z} \beta = P^T \beta
\]

by using \( \tilde{A}_I = A_I P \) and \( \tilde{\Sigma}^{z} = P^T \Sigma^{z} P \) in the last step. This concludes the proof.

#### B.2 Proof of Theorem 3: the minimax lower bounds for estimators of \( \beta \)

We first give two lemmas. Lemma 11 is proved in Section B.13 while Lemma 12 is proved in (Klopp et al., 2017, Lemma 16).

Let \( \mathbb{P}_{\beta} \) and \( \mathbb{P}_{\beta'} \) denote the joint distribution of \( (X_i, Y_i) \) for \( i = 1, \ldots, n \), parametrized by the same \( (A, \Sigma^{z}) \) but different \( \beta \) and \( \beta' \), respectively. Denote by \( \text{KL}(\mathbb{P}_{\beta}, \mathbb{P}_{\beta'}) \) the Kullback-Leibler divergence between these two distributions.

**Lemma 11.** Suppose that \( (X_i, Y_i) \) are i.i.d. Gaussian from model (1). Then, for any \( \beta, \beta' \in \mathbb{R}^K \),

\[
\frac{1}{n} \text{KL}(\mathbb{P}_{\beta}, \mathbb{P}_{\beta'}) \leq \frac{|\beta^T G^{-1} \beta - \beta'^T G^{-1} \beta'| + \| \Sigma^{z} - G^{-1} \|_{op} \| \beta - \beta' \|^2}{\sigma^2 + \min(\| \beta \|^2, \| \beta' \|^2) / \| G \|_{op}}.
\]
Lemma 12. Let $k \geq 2$ and $s \geq 1$ be integers, $s \leq k$. There exists a subset $S_0$ of the set of binary sequences $\{0,1\}^K$ such that

(i) $\log|S_0| \geq c_1 s \log(ek/s)$,

(ii) $c_2 s \leq \|a\|_0 \leq s$, for all $a \in S_0$, and all $\|a\|_0 = s$ for $a \in S_0$, if $s \leq k/2$,

(iii) $\|a - b\|^2 \geq c_3 s$, for all $a, b \in S_0$ and $a \neq b$,

where $c_j > 0$, $j = 1, 2, 3$ are absolute constants.

Proof of Theorem 3. Since

$$\sup_{(\beta,A,\Sigma^\beta) \in \mathcal{S}(R,m)} \mathbb{P}_A,\Sigma^\beta \left\{ \|\hat{\beta} - \beta\| \geq \epsilon' \left( 1 \vee \frac{R}{\sqrt{m}} \right) \cdot \sqrt{\frac{K}{n}} \right\} \geq \sup_{\|\beta\| \leq R} \mathbb{P}_A,\Sigma^\beta \left\{ \|\hat{\beta} - \beta\| \geq \epsilon' \left( 1 \vee \frac{R}{\sqrt{m}} \right) \cdot \sqrt{\frac{K}{n}} \right\},$$

for any fixed $A^*$ and $\Sigma_z^\ast$, we let $\mathbb{P}_\beta := \mathbb{P}_{A^*,\Sigma_z^\ast,\beta}$ for $A^*$ and $\Sigma_z^\ast$ defined below and aim to prove

$$\inf_{\beta} \sup_{\|\beta\| \leq R} \mathbb{P}_\beta \left\{ \|\hat{\beta} - \beta\| \geq \epsilon' \left( 1 \vee \frac{R}{\sqrt{m}} \right) \cdot \sqrt{\frac{K}{n}} \right\} \geq c''.$$

We choose $\Sigma_z^\ast$ such that $0 < C_{min} \leq \lambda_{min}(\Sigma_z^\ast) \leq \lambda_{max}(\Sigma_z^\ast) \leq C_{max} < \infty$ and

$$A^* := \begin{bmatrix} 1_m \otimes I_K & \\ 0 & \end{bmatrix}$$

with $\otimes$ denoting the kronecker product and $I_d$ denoting the vector in $\mathbb{R}^d$ with all ones.

We start by constructing a set of hypothesis $S$ for $\beta$. From Lemma 12 with $s = k = K - 1$, we can find a subset $S_0$ of the set of binary sequences $\{0,1\}^{K-1}$ such that

(i) $\log|S_0| \geq c_1 (K - 1),$

(ii) $c_2 (K - 1) \leq \|a\|_0 \leq (K - 1)$, for all $a \in S_0$.

(iii) $\|a - b\|^2 \geq c_3 (K - 1)$, for all $a, b \in S_0$ and $a \neq b$,

where $c_1, c_2, c_3 > 0$ are absolute constants. Let $v^{(0)} = (1,0,\ldots,0) \in \mathbb{R}^K$ and $v^{(j)} = (0,a) \in \mathbb{R}^K$ for all $a \in S_0$ so that $j \in \{1,\ldots,|S_0|\}$. We then define $\beta^{(0)} = (R,0,\ldots,0)$ and

$$\beta^{(j)} := \frac{R}{\sqrt{1 + \eta^2(K - 1)}} \left( v^{(0)} + \eta v^{(j)} \right) \text{ for all } j \in \{1,\ldots,|S_0|\},$$

with $\eta$ to be chosen later.

It is easy to verify that $\|\beta^{(j)}\| \leq R$ so that $(\beta^{(j)},\Sigma_z^\ast,A^*) \in \mathcal{S}(R,m)$ for $j \in \{0,1,\ldots,|S_0|\}$. Moreover, (iii) above implies that, for any $j, \ell \geq 1$ with $j \neq \ell$,

$$\|\beta^{(j)} - \beta^{(\ell)}\|^2 = \frac{R^2 \eta^2}{1 + \eta^2(K - 1)} \|v^{(j)} - v^{(\ell)}\|^2 \geq c_3 \frac{R^2 \eta^2}{1 + \eta^2(K - 1)}.$$

34
and (ii) above guarantees that, for any \( j \geq 1 \),
\[
\|\beta(j) - \beta(0)\|^2 \geq \frac{R^2(\sqrt{1 + \eta^2(K - 1)} - 1)^2}{1 + \eta^2(K - 1)} + \frac{R^2\eta^2}{1 + \eta^2(K - 1)}\|v(j)\|^2
\]
\[
\geq \frac{R^2\eta^2}{1 + \eta^2(K - 1)}\|v(j)\|^2
\]
\[
\geq c_2\frac{R^2\eta^2(K - 1)}{1 + \eta^2(K - 1)}.
\]  
(61)

On the other hand, for any \( j \in \{1, \ldots, |S_0|\} \), Lemma 11 implies
\[
\frac{1}{n}\text{KL}(P_{\beta(j)}, P_{\beta(0)}) \leq \frac{\|\beta(j)\|^2}{\sigma^2 + \min(\|\beta(j)\|^2, \|\beta(0)\|^2)/\|G\|_{\text{op}}}
\]
with \( G = \Omega + \tau^2 A^T A \). By using (59), the definition of \( \beta(0) \) and (ii), we further have
\[
\|\beta(j)\|^2 = R^2
\]
\[
\|\beta(j)\|^2 = \frac{R^2}{1 + \eta^2(K - 1)}\|v(0) + \eta v(j)\|^2 = \frac{R^2(1 + \eta^2\|v(0)\|^2)}{1 + \eta^2(K - 1)} \geq cR^2.
\]
Together with
\[
\|\beta(j) - \beta(0)\|^2 \leq \frac{R^2\eta^2(K - 1)}{4(1 + \eta^2(K - 1))} + \frac{R^2\eta^2}{1 + \eta^2(K - 1)}\|v(j)\|^2 \leq \frac{5R^2\eta^2(K - 1)}{4(1 + \eta^2(K - 1))},
\]  
from (61) and the fact that \( f(x) = \sqrt{x} \) is concave for \( x > 0 \), we obtain
\[
\text{KL}(P_{\beta(j)}, P_{\beta(0)}) \leq \frac{5}{4} \cdot \frac{nR^2\eta^2(K - 1)}{1 + \eta^2(K - 1)} \cdot \frac{\|G^{-1}\|_{\text{op}} + \|\Sigma^z - G^{-1}\|_{\text{op}}}{\sigma^2 + \min(\|\beta(j)\|^2, \|\beta(0)\|^2)/\|G\|_{\text{op}}} \leq 3n\eta^2(K - 1) \cdot \frac{R^2C_{\text{max}}}{\sigma^2 + cR^2/\|G\|_{\text{op}}},
\]
where in the second line we use
\[
\|G^{-1}\|_{\text{op}} \leq \|\Sigma^z\|_{\text{op}} \left\|\left[I_K + (\Sigma^z)^{1/2} A^T \Omega^w A(\Sigma^z)^{1/2}\right]^{-1}\right\|_{\text{op}} \leq \|\Sigma^z\|_{\text{op}}.
\]  
35
Further note that
\[ \|G\|_{op} \leq \|\Omega\|_{op} + \tau^{-2}\|A^T A\|_{op} \leq C^{-1}_{\min} + m/\tau^2 \leq c'm/\tau^2. \] (63)

Choosing
\[ \eta^2 = c \cdot \frac{\sigma^2 + c\tau^2 R^2/(c'm)}{nR^2 C_{\max}} \] (64)
yields
\[ KL(\mathbb{P}_{\beta(j)}, \mathbb{P}_{\beta(\ell)}) \leq c \log |S_0| \]
for any \( j \geq 1 \), and
\[ \|\beta(j) - \beta(\ell)\|^2 \geq c \left( \frac{\sigma^2}{C_{\max}} \vee \frac{\tau^2 R^2}{c'C_{\max} m} \right) \frac{(K - 1)}{n} \cdot \frac{1}{1 + \eta^2(K - 1)} \]
for any \( j \neq \ell \), from (60) and (62). Since condition \( K \leq c(R^2 \vee m)n \) guarantees that \( \eta^2(K - 1) \leq c = c(\bar{c}) \), invoking Theorem 2.5 in Tsybakov (2009) concludes
\[ \inf_{\beta} \sup_{||\beta|| \leq R} \mathbb{P}_{\beta} \left\{ \|\hat{\beta} - \beta\| \geq c \left( 1 \vee R \frac{\sqrt{K - 1}}{n} \right) \right\} \geq c', \]
which completes the proof. \( \square \)

### B.3 Notation and preliminary lemmas

In the sequel, we let \( X, Z, A, W \) and \( \varepsilon \) be the corresponding parts of the following model:
\[ Y = Z\beta + \varepsilon, \quad X = ZA^T + W, \] (65)
where \( X = (x_1, \ldots, x_n)^T = (X_1, \ldots, X_p), Z = (z_1, \ldots, z_n)^T = (Z_1, \ldots, Z_K) \) and \( W = (W_1, \ldots, W_p) \).

Define the event
\[ \mathcal{E} := \left\{ \max_{k \in [K]} \frac{1}{n} \sum_{i=1}^{n} Z^2_{ik} \leq B_z \right\} \bigcap \left\{ \max_{1 \leq j < \ell \leq p} |\bar{\Sigma}_{j\ell} - \Sigma_{j\ell}| \leq \delta \right\} \] (66)
with \( B_z \) defined in Assumption 2 and \( \delta \asymp \sqrt{\log(p \vee n)/n} \). Note that, under \( \log p = o(n) \), Lemma 1 in Bien et al. (2016) together with Lemmas 13 and 15 below guarantee that \( \mathbb{P}(\mathcal{E}) \geq 1 - (p \vee n)^{-c} \) for some constant \( c > 0 \). Moreover, Bing et al. (2017) guarantees that, on the event \( \mathcal{E} \), Theorem 2 holds.

Let \( \hat{I} := \{\hat{I}_1, \ldots, \hat{I}_K\} \) be the output from Algorithm 1. For simplicity, we assume the identity group permutation \( \pi \) in Theorem 2 and \( A_{ik} = 1 \) for any \( i \in I_k \) and \( k \in [K] \) (for the general case, the signed permutation \( P \) can be traced throughout the proofs). We write
\[ \hat{X} := X_{\hat{I}}\hat{A}_{\hat{I}}(\hat{A}_{\hat{I}}^T \hat{A}_{\hat{I}})^{-1}, \quad \hat{Z} := ZA^T_{\hat{I}}\hat{A}_{\hat{I}}(\hat{A}_{\hat{I}}^T \hat{A}_{\hat{I}})^{-1}, \quad \hat{W} := W_{\hat{I}}\hat{A}_{\hat{I}}(\hat{A}_{\hat{I}}^T \hat{A}_{\hat{I}})^{-1}, \] (67)
such that \( \hat{X} = \hat{Z} + \hat{W} \). For all \( k \in [K] \), let \( \hat{m}_k := |\hat{I}_k|, \| \hat{m}_k := |I_k|, m := \min_k m_k \) and define
\[ \hat{\eta}_k = \frac{1}{\hat{m}_k} \sum_{i \in \hat{I}_k} W_{ti}, \quad \eta_k = \frac{1}{m_k} \sum_{i \in I_k} W_{ti}, \quad \tau^2_k := \frac{1}{m_k^2} \sum_{i \in I_k} \tau^2_{i}. \] (68)

36
We further let $L_k := \hat{I}_k \setminus I_k \subseteq J_k$ such that $\hat{m}_k = m_k + |L_k|$. Finally, the following is repeatedly used in the proof, which follows directly from the definition of $J_k^1$:

$$\max_{k \in [K], j \in J_k^1} \| A_j - e_k \|_1 \leq 8\delta/\nu. \tag{69}$$

with $\nu$ defined in Assumption 1. We use $c, c'$ to denote positive absolute constants which may vary line by line.

We next give three preliminary lemmas that will be used throughout the proofs that follow, specifically, in the proofs of Theorems 4 and 5 and Proposition 7. Their proofs are deferred to Section B.12.

**Lemma 13.** Suppose Assumptions 2 and 4 hold.

(1) For any fixed $v$, $\langle Z, v \rangle$ is $(\gamma_z \sqrt{v^T \Sigma v})$-subgaussian, hence $(\|v\| \gamma_z \sqrt{C_{\text{max}}})$-subgaussian. In particular, $Z_k$ is $\gamma'_z$-subgaussian with $\gamma'_z := \gamma_z \sqrt{B_z}$ for any $k \in [K]$.

(2) $X_j$ is $(\gamma'_z + \gamma_w)$-subgaussian for any $j \in [p]$.

**Lemma 14.** Suppose Assumption 2 holds and let $\eta_{tk}$ be defined as in (68). Then $\eta_{tk}$ is $(\gamma_w / \sqrt{m})$-subgaussian. Moreover, for any vector $v \in \mathbb{R}^K$ with $\|v\| \leq 1$, $\sum_{k \in [K]} v_k \eta_{tk}$ is also $(\gamma_w / \sqrt{m})$-subgaussian.

**Lemma 15.** Let $\{X_t\}_{t=1}^n$ and $\{Y_t\}_{t=1}^n$ be any two sequences, each with zero mean independent $\gamma_x$-subgaussian and $\gamma_y$-subgaussian elements. Then, for some constants $c, c' > 0$, we have

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{t=1}^n (X_t Y_t - \mathbb{E}[X_t Y_t]) \leq c \gamma_x \gamma_y t \right\} \geq 1 - 2 \exp \{ -c' \min (t^2, t) n \} .$$

In particular, when $\log p = o(n)$, one has

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{t=1}^n (x_t y_t - \mathbb{E}[x_t y_t]) \leq c \gamma_x \gamma_y \sqrt{\frac{\log p}{n}} \right\} \geq 1 - 2 p^{-c'} .$$

**B.4 Proof of Theorem 4 (convergence rate of $\hat{\beta}$) and statement (38) in Section 3**

We work on the event $\mathcal{E}$, defined in (66), intersected with the event

$$\mathcal{E}_1 := \left\{ \| \hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}} \|_{op} \leq c (C_{\text{max}} \vee \rho \sqrt{K}) \sqrt{\frac{K \log (p \vee n)}{n}} \right\} \tag{70}$$

where $c > 0$ is a constant such that by Lemma 21 in Section B.7, $\mathbb{P}(\mathcal{E}_1) \geq 1 - (p \vee n)^{-c'}$. Since Assumption 3 and $K \log (p \vee n) = O(n)$ imply

$$c''(1 \vee \rho^2 K) K \log (p \vee n) \leq n \tag{71}$$

for a sufficiently large $c'' > 0$, by using Weyl's inequality, we have

$$\lambda_{\min}(\hat{\Sigma}^{\frac{1}{2}}) \geq C_{\min} - \| \hat{\Sigma}^{\frac{1}{2}} - \Sigma^{\frac{1}{2}} \|_{op} \geq (1 - c_0) C_{\min} \tag{72}$$
for some constant $c_0 \in (0, 1)$. Hence $\hat{\Sigma}^z$ is invertible.

Recall that $\hat{\Sigma}^z \hat{\beta} = \hat{h}$. It follows from (65) and (13) that

$$
\hat{\Sigma}^z (\hat{\beta} - \beta) = \frac{1}{n} (\hat{A}^T \hat{A})^{-1} \hat{A}^T X^T (Z \beta + \varepsilon) - \hat{\Sigma}^z \beta
$$

$$
= \frac{1}{n} \bar{X}^T Z \beta - \hat{\Sigma}^z \beta + \frac{1}{n} \bar{X}^T \varepsilon. \tag{67}
$$

We observe that the definition of $\hat{\Sigma}^z$ in (53) yields

$$
\hat{\Sigma}^z_{aa} = \frac{1}{n} \bar{X}_a^T \bar{X}_a - d_a, \forall a \in [K]; \quad \hat{\Sigma}^z_{ab} = \frac{1}{n} \bar{X}_a^T \bar{X}_b, \forall a \neq b \in [K] \tag{73}
$$

with $\bar{X}_a$ being the $a$th column of $\bar{X}$, and

$$
d_a = \frac{1}{m_a} \sum_{i \in I_a} \frac{1}{n} X_i^T X_i - \frac{1}{m_a (m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} X_i^T X_j. \tag{74}
$$

Let $D = \text{diag}(d_1, \ldots, d_K)$ so that $\hat{\Sigma}^z = n^{-1} \bar{X}^T \bar{X} - D$. Plugging it into $\hat{\Sigma}^z (\hat{\beta} - \beta)$ and using $\bar{X} = \bar{Z} + \bar{W}$ twice yield

$$
\hat{\Sigma}^z (\hat{\beta} - \beta) = \frac{1}{n} \bar{X}^T (Z - \bar{X}) \beta + \frac{1}{n} \bar{X}^T \varepsilon + D \beta
$$

$$
= \frac{1}{n} \bar{X}^T (Z - \bar{Z}) \beta - \frac{1}{n} \bar{X}^T \bar{W} \beta + \frac{1}{n} \bar{X}^T \varepsilon + D \beta;
$$

$$
= \frac{1}{n} \bar{X}^T (Z - \bar{Z}) \beta - \frac{1}{n} \bar{Z}^T \bar{W} \beta + \frac{1}{n} \bar{X}^T \varepsilon + \left(D - \frac{1}{n} \bar{W}^T \bar{W}\right) \beta.
$$

We now bound the sup-norm of each term separately. Recall that $\hat{K} = K$ on the event $E$.

$$
\left| \frac{1}{n} \bar{X}_k^T \varepsilon \right| = \left| \frac{1}{m_k} \sum_{i \in I_k} \frac{1}{n} X_i^T \varepsilon \right| \leq \max_{i \in I_k} \frac{1}{n} \sum_{t=1}^{n} X_{ti} \varepsilon_t \right|.
$$

Note that $X_{ti}$ is $\gamma_x$-subgaussian from part (2) of Lemma 13. From the result of Theorem 2 that $I_k \subseteq I_k \cup J_k^c$, invoking Lemma 15 and taking the union bounds over $i \in I_k \cup J_k^c$ and $k \in [K]$ give

$$
P \left\{ \frac{1}{n} \| \bar{X}^T \varepsilon \|_\infty \leq c \left( \log (p \vee n) \right) \right\} \geq 1 - (p \vee n)^{-c}. \tag{75}
$$

The upper bounds of $\| \bar{Z}^T \bar{W} \beta \|_\infty$ and $n^{-1} \| \bar{X}^T (Z - \bar{Z}) \beta \|_\infty$ are given, respectively, in Lemma 17 by taking $v = \beta$ and Lemma 19 in Section B.7. Moreover, choosing $u = e_k$, $v = \beta$ and taking the union bounds over $k \in [K]$ in Lemma 18 in Section B.7 yield

$$
P \left\{ \| (D - n^{-1} \bar{W}^T \bar{W}) \beta \|_\infty \leq c \left( \frac{\| \beta \|}{\sqrt{m}} + \rho \| \beta \|_1 \right) \sqrt{\frac{\log (p \vee n)}{n}} \right\} \geq 1 - (p \vee n)^{-c}. \tag{76}
$$

Combining the results of the three terms gives

$$
P \left\{ \| \hat{\Sigma}^z (\hat{\beta} - \beta) \| \leq c \left( 1 \vee \frac{\| \beta \|}{\sqrt{m}} \vee \rho \| \beta \|_1 \right) \sqrt{\frac{K \log (p \vee n)}{n}} \right\} \geq 1 - (p \vee n)^{-c}. \tag{77}
$$
Since \( \| \Sigma^z(\hat{\beta} - \beta) \| \geq \lambda_{\min}(\Sigma^z)\| \hat{\beta} - \beta \| \geq (1 - c_0)C_{\min}\| \hat{\beta} - \beta \| \) from (72), we conclude
\[
\min_{P \in \mathcal{H}_K} \| \hat{\beta} - P\beta \| \lesssim C_{\min}^{-1} \left( 1 \lor \frac{\| \beta \|}{\sqrt{m}} \lor \rho \| \beta \|_1 \right) \sqrt{\frac{K \log(p \lor n)}{n}}
\]
with probability greater than \( 1 - (p \lor n)^{-c} \). Using \( \| \beta \|_1 \leq \sqrt{K}\| \beta \| \) and \( \rho = O(1/\sqrt{mK}) \) from Assumption 3 finishes the proof of (23).

We proceed to show (38). By writing \( \Delta = \hat{\beta} - \beta \), we have
\[
\frac{1}{n} \| Z\Delta \|^2 \leq \left| \Delta^T \left( \frac{1}{n} Z^T Z - \hat{\Sigma}^z \right) \Delta \right| + \left| \Delta^T \hat{\Sigma}^z \Delta \right|
\]
\[
\leq \| \Delta \|^2 \left( \| n^{-1} Z^T Z - \Sigma^z \|_{\text{op}} + \| \hat{\Sigma}^z - \Sigma^z \|_{\text{op}} \right) + \| \Delta \| \| \hat{\Sigma}^z \|.
\]
Using Lemma 21 in Section B.7, the rate of \( \| \hat{\beta} - \beta \| \), (77) and \( K \log(p \lor n) = O(n) \) completes the proof.

\[\square\]

### B.5 Proof of (26): convergence rate of \( \hat{\beta}_d \)

We recall the definition of the \( \ell_q \)-sensitivity
\[
\kappa_q(\Sigma^z, s) := \inf_{|S| \leq s} \inf_{v \in C_S} \frac{\| \Sigma^z v \|_\infty}{\| v \|_q}.
\]
The first inf is taken over index sets \( S \subseteq \{1, \ldots, K\} \) of size \( |S| \) at most \( s \), while the second inf is taken over vectors \( v \) in the set \( C_S := \{ v \in \mathbb{R}^K : \| v_{S^c} \|_1 \leq \| v_S \|_1 \} \). We first give the precise statement of (26).

**Theorem 16.** Assume Assumptions 1 – 2 hold. Let \( s = |\text{supp}(\beta)| = \| \beta \|_0 \). Then \( \hat{\beta}_d \) constructed as (25) by choosing \( \mu \asymp \lambda \asymp \sqrt{\log(p \lor n)/n} \) satisfies:
\[
\min_{P \in \mathcal{H}_K} \left\| \hat{\beta}_d - P\beta \right\|_q \lesssim [\kappa_q(\Sigma^z, s)]^{-1} (1 \lor \| \beta \|_1) \sqrt{\frac{s \log(p \lor n)}{n}}
\]
with probability greater than \( 1 - (p \lor n)^{-c} \) for some constant \( c > 0 \).

**Proof.** We first show that the true \( \beta \) lies in the feasible set with high probability by verifying that
\[
\| \Sigma^z \beta - \tilde{h} \|_\infty \lesssim (1 \lor \| \beta \|_1) \sqrt{\log(p \lor n)/n}
\]
holds with high probability. This follows from the proof of Theorem 4 by observing that
\[
\tilde{h} - \Sigma^z \beta = \frac{1}{n} X^T \varepsilon - \frac{1}{n} Z^T W \beta + \frac{1}{n} X^T (Z - \bar{Z}) \beta + \left( D - \frac{1}{n} W^T W \right) \beta
\]
and, to bound the sup-norm of the right hand side with high probability, using (75), Lemmas 17 and 19 in Section B.7 with \( v = \beta \), and (76) on each of the four terms respectively. This gives
\[
P \left\{ \| \Sigma^z \beta - \tilde{h} \|_\infty \leq c (1 \lor \frac{\| \beta \|}{\sqrt{m}} \lor \rho \| \beta \|_1) \sqrt{\frac{\log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c},
\]
\[
\hat{\beta}_d
\]
\[
\frac{1}{n} \| Z\Delta \|^2 \leq \left| \Delta^T \left( \frac{1}{n} Z^T Z - \hat{\Sigma}^z \right) \Delta \right| + \left| \Delta^T \hat{\Sigma}^z \Delta \right|
\]
\[
\leq \| \Delta \|^2 \left( \| n^{-1} Z^T Z - \Sigma^z \|_{\text{op}} + \| \hat{\Sigma}^z - \Sigma^z \|_{\text{op}} \right) + \| \Delta \| \| \hat{\Sigma}^z \|.
\]
Using Lemma 21 in Section B.7, the rate of \( \| \hat{\beta} - \beta \| \), (77) and \( K \log(p \lor n) = O(n) \) completes the proof.
and (80) follows from using $\rho \leq 1$ and $\|\beta\| \leq \|\beta\|_1$. On the event that $\beta$ is feasible, we have $\|\hat{\beta}_d\|_1 \leq \|\beta\|_1$ by construction, whence $\|((\hat{\beta}_d - \beta)_{S^c})_1 \leq \|(\hat{\beta}_d - \beta)_{S^c}\|_1$ for $S = \text{supp}(\beta)$ and we further obtain,

\[
\kappa_q(\Sigma^z, s)\|\hat{\beta}_d - \beta\|_q \leq \|\Sigma^z \hat{\beta}_d - \Sigma^z \beta\|_\infty \\
\leq \|\Sigma^z \hat{\beta}_d - \Sigma^z \beta - \hat{h}\|_\infty \\
\leq \|\Sigma^z - \Sigma^z\|_\infty \|\hat{\beta}_d - \beta\|_1 + 2\lambda + 2\mu \|\beta\|_1 \\
\leq 2\|\beta\|_1 \|\Sigma^z - \Sigma^z\|_\infty + 2\lambda + 2\mu \|\beta\|_1.
\]

Invoking Lemma 20 in Section B.7 with $u = v = e_k$ and taking the union bounds over $k \in [K]$, we conclude that

\[
\kappa_q(\Sigma^z, s)\|\hat{\beta}_d - \beta\|_q \lesssim (1 \lor \|\beta\|_1) \sqrt{\log(p \lor n)/n}
\]

with probability greater than $1 - (p \lor n)^{-c}$.

\[
\square
\]

### B.6 Proof of Theorem 5: asymptotic normality of $\hat{\beta}$

We work on the event $E \cap E_1$ defined in (66) and (70), respectively. Thus $\hat{\Sigma}^z$ is invertible, as shown in the proof of Theorem 4. We further have $\tilde{A}_f = A_f$. From (14), we have

\[
\hat{\beta} = (\hat{\Sigma}^z)^{-1} \tilde{h} = (\hat{\Sigma}^z)^{-1} \left( \tilde{A}_f^T \tilde{A}_f \right)^{-1} \tilde{A}_f^T X^T Y \overset{(67)}{=} (\hat{\Sigma}^z)^{-1} \frac{1}{n} X^T Y.
\]

Write $\tilde{X} := X_f A_f (A_f^T A_f)^{-1}$ and $\tilde{W} := W_f A_f (A_f^T A_f)^{-1}$. Analogous to (73) and (74), further write $\Sigma^z = \frac{1}{n} \tilde{X}^T \tilde{X} - \tilde{D}$ with $\tilde{D} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_K)$ and

\[
\begin{align*}
\tilde{d}_a &= \frac{1}{m_a} \sum_{i \in I_a} \frac{1}{n} X_i^T X_i - \frac{1}{m_a(m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} X_i^T X_j, \quad \text{for any } a \in [K].
\end{align*}
\]

Notice that model (1) implies $Y = Z \beta + \varepsilon$ and $\tilde{X} = Z + \tilde{W}$. Plugging these into the above expression of $\hat{\beta}$ and adding and subtracting terms give

\[
\begin{align*}
\hat{\beta} - \beta &= (\hat{\Sigma}^z)^{-1} \left[ \frac{1}{n} \tilde{X}^T Z \beta + \frac{1}{n} \tilde{X}^T \varepsilon - \Sigma^z \beta + \frac{1}{n} (\tilde{X} - \tilde{X})^T Y - (\hat{\Sigma}^z - \Sigma^z) \beta \right] \\
&= (\hat{\Sigma}^z)^{-1} \left[ \frac{1}{n} \tilde{X}^T (\tilde{X} - \tilde{W}) \beta + \frac{1}{n} (Z + \tilde{W})^T \varepsilon - (\hat{\Sigma}^z - \Sigma^z) \beta \right] \\
&\quad + (\hat{\Sigma}^z)^{-1} \left[ \frac{1}{n} (\tilde{X} - \tilde{X})^T Y - (\hat{\Sigma}^z - \Sigma^z) \beta \right] \\
&\overset{\text{Rem}_1}{=} (\hat{\Sigma}^z)^{-1} \left[ \Sigma^w \beta + \frac{1}{n} (Z + \tilde{W})^T \varepsilon - \frac{1}{n} Z^T \tilde{W} \beta \right] + \text{Rem}_1 \\
&= (\Sigma^z)^{-1} \left[ \Sigma^w \beta + \frac{1}{n} (Z + \tilde{W})^T \varepsilon - \frac{1}{n} Z^T \tilde{W} \beta \right] + \text{Rem}_1 + \text{Rem}_2,
\end{align*}
\]

with $\Sigma^w = n^{-1} \tilde{X}^T \tilde{X} - \Sigma^z - n^{-1} \tilde{W}^T \tilde{W}$ and

\[
\begin{align*}
\text{Rem}_2 &= \left[ (\hat{\Sigma}^z)^{-1} - (\Sigma^z)^{-1} \right] \left[ \Sigma^w \beta + \frac{1}{n} (Z + \tilde{W})^T \varepsilon - \frac{1}{n} Z^T \tilde{W} \beta \right].
\end{align*}
\]

40
We first simplify the first term in (81). Using (73) with $\tilde{X} = \bar{X}$ and $\tilde{I} = I$, we have

$$\Sigma^w_{ab} = -\frac{1}{n} W^T_a W_b = -\frac{1}{m_a m_b} \sum_{i \in I_a, j \in I_b} \frac{1}{n} W_i^T W_j,$$

(82)

for any $a \neq b \in [K]$, and, by using $X_i = Z_a + W_i$ for any $i \in I_a$ and $a \in [K]$,

$$\Sigma^w_{aa} = \frac{1}{m_a^2} \sum_{i,j \in I_a} \frac{1}{n} X_i^T X_j - \frac{1}{m_a(m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} X_i^T X_j - \frac{1}{m_a^2} \sum_{i,j \in I_a} \frac{1}{n} W_i^T W_j$$

$$= \frac{1}{m_a^2} \sum_{i,j \in I_a} \frac{1}{n} (Z_a^T Z_a + Z_a^T W_j + Z_a^T W_i + W_i^T W_j) - \frac{1}{m_a(m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} (Z_a^T W_j + Z_a^T W_i + W_i^T W_j)$$

$$- \frac{1}{m_a(m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} W_i^T W_j$$

$$= -\frac{1}{m_a(m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} W_i^T W_j,$$

for any $a \in [K]$. Therefore, for any $k \in [K]$, it follows that

$$\sqrt{n} (\tilde{\beta}_k - \beta_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Omega^T_k U_i + \sqrt{n} [Rem_1)_k + \sqrt{n} [Rem_2]_k$$

(83)

where we write $U_i := u^T \beta + (Z_t + \tilde{W}_t) \varepsilon_t - Z_t (\tilde{W}_t', \beta) \in \mathbb{R}^K$ with

$$u^T_{aa} = -\frac{1}{m_a(m_a - 1)} \sum_{i \neq j \in I_a} W_i^T W_j, \quad \text{for any } a \in [K]$$

(84)

$$u^T_{ab} = -\frac{1}{m_a m_b} \sum_{i \in I_a, j \in I_b} W_i^T W_j, \quad \text{for any } a \neq b \in [K].$$

From (83), we aim to apply the Liapunov central limit theorem (see for instance, ?) to the first term and control the two remainder terms afterwards.

Note that $\Omega^T_k U_t$ is independent across $t$ and $\mathbb{E}[U_t] = 0$ from our model specifications. Moreover,

$$\mathbb{E}[U^2_{tk}] = \mathbb{E} \left[ (u^T_t, \beta)^2 + (Z_t + \tilde{W}_{tk})^2 \varepsilon_t^2 + Z_{tk}^2 (\tilde{W}_t', \beta)^2 \right], \quad \forall k \in [K]$$

by using the independence among $\varepsilon$, $Z$ and $W$ to cancel the cross terms. To calculate the three expectations on the right-hand-side, recall that $W_{ti} \perp W_{tj}$ for any $i \neq j \in [p]$. It implies $\tilde{W}_{ta} \perp \tilde{W}_{tb}$ for any $a \neq b \in [K]$ and

$$\mathbb{E} [\tilde{W}^2_{ta}] = \mathbb{E} \left[ \frac{1}{m_a^2} \left( \sum_{i \in I_a} W_{ti} \right)^2 \right] = \mathbb{E} \left[ \frac{1}{m_a^2} \sum_{i \in I_a} W^2_{ti} \right], \quad \text{(68)}$$

$$= \pi_a^2.$$
for any $a \in [K]$. It is straightforward to verify that

$$
\mathbb{E} \left[ (Z_{tk} + \bar{W}_{tk})^2 \epsilon_t^2 \right] = \sigma^2 \left( \sum \tau_i^2 \right) = \sigma^2 \left( \sum \tau_k^2 \right),
$$

(85)

$$
\mathbb{E} \left[ Z_{tk}^2 (W_t, \beta)^2 \right] = \sum \frac{\beta^2}{m^2} \sum \tau_i^2 = \sum \beta^2 \tau_k^2.
$$

(86)

From (84), using the independence of the entries of $W$ again gives $\mathbb{E}[u_{tb} u_{tc}] = 0$, for any $b \neq c \in [K]$ and $\mathbb{E}[W_{ti} u_{tb}] = 0$, for any $a, b \in [K], t \in [n]$ and $i \in [p]$. This further gives

$$
\mathbb{E} \left[ (u_{tk}, \beta)^2 \right] = \sum \beta^2 \mathbb{E} \left[ (u_{ka})^2 \right]
$$

$$
= \beta^2 \mathbb{E} \left[ \sum \frac{m^2}{m^2 (m - 1)^2} \right] + \sum \beta^2 \mathbb{E} \left[ \sum \frac{W_t^2 W_i^2}{m^2 m^2} \right]
$$

$$
= \frac{\beta^2}{m^2 (m - 1)^2} \sum \tau_i \tau_j + \sum \beta^2 \tau_k \tau_a
$$

(87)

Collecting (85) – (87) yields

$$
\mathbb{E}[U_{tk}]
$$

$$
= \sum \frac{\beta^2}{m^2} \mathbb{E} \left[ (u_{ka})^2 \right]
$$

$$
= \sum \frac{\beta^2}{m^2} \left( \sigma^2 + \sum \beta^2 \tau_k \right) + \sigma^2 \left( \sum \tau_i \tau_j \right) + \sum \beta^2 \tau_k \left( \sum \frac{\beta^2}{m^2 (m - 1)^2} \right)
$$

$$
= \sum \frac{\beta^2}{m^2} \left( \sigma^2 + \sum \beta^2 \tau_a \right) + \sigma^2 \left( \sum \beta^2 \tau_a \right) + \sum \beta^2 \tau_k \left( \sum \frac{\beta^2}{m^2 (m - 1)^2} \right)
$$

$$
= \sum \frac{\beta^2}{m^2} \left( \sigma^2 + \sum \beta^2 \tau_a \right)
$$

for any $k \in [K]$. Similarly, one can derive, for any $k \neq \ell \in [K]$,

$$
\mathbb{E}[U_{tk} U_{\ell \ell}] = \mathbb{E} \left[ (Z_{tk} + \bar{W}_{tk})(Z_{\ell \ell} + \bar{W}_{\ell \ell}) \epsilon_t^2 \right] - \mathbb{E}[Z_{tk} Z_{\ell \ell} (\bar{W}_{t}, \beta)^2]
$$

$$
= \sum \frac{\beta^2}{m^2} \left( \sigma^2 + \sum \beta^2 \tau_a \right)
$$

42
Let $D = \text{diag}(D_1, \ldots, D_K)$ and recall $\Omega = (\Sigma^2)^{-1}$. It follows that
\[
\frac{1}{n} \sum_{t=1}^{n} \Omega_k^T \mathbb{E}[U_t U_t^T] \Omega_k = \Omega_k^T \left( \sum_{a=1}^{K} \beta_a^2 \tau_a^2 \right) \Omega_k + \sum_{a=1}^{K} \Omega_{ka}^2 \Omega_{ka} D_a
\]
\[
= \left( \sum_{a=1}^{K} \beta_a^2 \tau_a^2 \right) \Omega_{kk} + \sum_{a=1}^{K} \Omega_{ka}^2 \Omega_{ka}
\]
\[
+ \sum_{a=1}^{K} \Omega_{ka}^2 \beta_a^2 \left[ \frac{\sum_{i \neq j \in I_a} \tau_i^2 \tau_j^2}{m_a^2 (m_a - 1)^2} - \tau_i^2 \right].
\]

When $\tau_i^2 = \tau^2$ and $m_k = m$ for all $i \in I$ and $k \in [K]$, the above is equal to $Q_{kk}$ as desired.

It remains to check the Liapunov condition, for which it suffices to show
\[
\lim_{n \to \infty} \sum_{t=1}^{n} \mathbb{E}[\|\Omega_k^T U_t|\|^3] / (nQ_{kk})^{3/2} = 0. \quad (88)
\]

Recall that
\[
\Omega_k^T U_t = \sum_{a=1}^{K} \Omega_{ka} \{ (u_{a, \cdot}, \beta) + W_{ta} \varepsilon_t + Z_{ta} (\varepsilon_t - (W_t, \beta)) \}.
\]

By using the inequality $|x + y|^3 \leq 4(|x|^3 + |y|^3)$ twice, we immediately obtain
\[
\mathbb{E}[\|\Omega_k^T U_t|^3]
\leq 16 \left\{ \mathbb{E} \left[ \|\Omega_k^T (u^t, \beta)\|^3 \right] + \mathbb{E} \left[ \|\Omega_k^T W_t, \varepsilon_t\|^3 \right] + \mathbb{E} \left[ \|\Omega_k^T Z_t (\varepsilon_t - (W_t, \beta))\|^3 \right] \right\}.
\]

We upper bound the first two terms via their respective $\| \cdot \|_{\psi_1}$ norms. Since (84) implies
\[
\langle u^t, \beta \rangle = -\frac{\beta_a}{m(m-1)} \sum_{i \neq j \in I_a} W_{ti} W_{tj} - \sum_{b \neq a} \beta_b W_{ta} W_{tb}
\]
\[
= -\frac{\sum_{b=1}^{K} \beta_b W_{ta} W_{tb} + \beta_a W_{ta}^2}{m(m-1)} - \frac{\beta_a}{m^2 (m-1)} \sum_{i \neq j \in I_a} W_{ti} W_{tj}
\]
\[
= -\bar{W}_{ta} \sum_{b=1}^{K} \beta_b W_{tb} - \frac{\beta_a}{m^2 (m-1)} \sum_{i \neq j \in I_a} W_{ti} W_{tj} + \frac{\beta_a}{m^2} \sum_{i \in I_a} W_{ti}^2, \quad (89)
\]

we obtain
\[
\| \Omega_k^T \langle u^t, \beta \rangle \|_{\psi_1}
\leq \| (\Omega_k^T \bar{W}_t) (\beta^T \bar{W}_t) \|_{\psi_1} + \frac{|\Omega_k^T \beta|}{m} \left( \max_{a \in [K]} \| W_{ta} W_{tj} \|_{\psi_1} + \max_{a \in [K]} \| W_{ti} \|_{\psi_1} \right).
\]

43
Note that $\bar{W}_t$ is independent with $\bar{W}_b$ for any $b \neq a$. Also note that $\|\bar{W}_t\|_{\psi_2} \leq c(g)\sqrt{m}$ for some constant $c > 0$, from the subgaussianity of $W$. We then have

$$
\|\Omega^T_k \bar{W}_t\|_{\psi_2} \leq c\|\Omega_k\| \frac{\gamma}{\sqrt{m}}, \quad \|\beta^T \bar{W}_t\|_{\psi_2} \leq c\|\beta\| \frac{\gamma}{\sqrt{m}}.
$$

By triangle inequality and using $\|XY\|_{\psi_2} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$ for any two subgaussian random variables $X$ and $Y$, we conclude

$$
\|\Omega^T_k \langle u^t, \beta \rangle\|_{\psi_1} \leq c\|\Omega_k\| \frac{\|\beta\|}{m},
$$

from which, the definition of $\| \cdot \|_{\psi_1}$ further yields

$$
\mathbb{E} \left[ |\Omega^T_k \langle u^t, \beta \rangle|^3 \right] \leq c \frac{\|\Omega_k\|^3 \|\beta\|^3}{m^3}.
$$

Similarly, $\|\Omega^T_k \bar{W}_t \varepsilon_t\|_{\psi_1}$ can be upper bounded by

$$
\|\Omega^T_k \bar{W}_t\|_{\psi_2} \|\varepsilon_t\|_{\psi_2} \leq c\|\Omega_k\| \frac{\gamma}{\sqrt{m}},
$$

such that

$$
\mathbb{E} \left[ |\Omega^T_k \bar{W}_t \varepsilon_t|^3 \right] \leq c \frac{\|\Omega_k\|^3}{m^3/2}.
$$

Finally, we bound the third term. The independence of $Z$, $W$ and $\varepsilon$ guarantees

$$
\mathbb{E} \left[ |\Omega^T_k \zeta_t (\varepsilon_t - \langle \bar{W}_t, \beta \rangle)|^3 \right] \leq \mathbb{E} \left[ |\Omega^T_k \zeta_t|^3 \right] \mathbb{E} \left[ |\varepsilon_t - \langle \bar{W}_t, \beta \rangle|^3 \right].
$$

Note that $\|\varepsilon_t - \langle \bar{W}_t, \beta \rangle\|_{\psi_2} \leq c(\gamma \varepsilon + \|\beta\| \gamma\sqrt{m})$ and part (1) of Lemma 13 implies that $\langle \Omega_k, \zeta_t \rangle$ is $(\gamma \varepsilon \sqrt{\Omega_{kk}})$-subgaussian. The definition of $\| \cdot \|_{\psi_2}$ implies

$$
\mathbb{E} \left[ |\Omega^T_k \zeta_t|^3 \right] \mathbb{E} \left[ |\varepsilon_t - \langle \bar{W}_t, \beta \rangle|^3 \right] \leq c \left( 1 + \frac{\|\beta\|}{\sqrt{m}} \right)^3 \Omega_{kk}^{3/2}.
$$

Collecting (90) - (92) yields

$$
n^{-3/2} \sum_{t=1}^n \mathbb{E} \left[ |\Omega^T_k U_t|^3 \right] \leq c \frac{\Omega_{kk} \vee \|\Omega_k\|^2}{m} \left( 1 + \frac{\|\beta\|^2}{m} \right)^{3/2}.
$$

Recall that

$$
Q_{kk}^{3/2} \geq \left( \sigma^2 + \frac{\tau^2 \|\beta\|^2}{m} \right)^{3/2} \left( \Omega_{kk} \vee \frac{\tau^2 \|\Omega_k\|^2}{m} \right)^{3/2}.
$$

We thus conclude that

$$
\lim_{n \to \infty} \frac{\sum_{t=1}^n \mathbb{E} \left[ |\Omega^T_k U_t|^3 \right]}{(n Q_{kk})^{3/2}} \leq \lim_{n \to \infty} \frac{1}{\sqrt{n}} = 0.
$$

Invoking Liapunov central limit theorem, we have

$$
\frac{1}{\sqrt{n}} \sum_{t=1}^n \Omega^T_k U_t \to N(0, Q_{kk}), \quad \text{as } n \to \infty.
$$
From (83), it remains to control the two remainder terms by showing

\[ \sqrt{n}\|\text{Rem}_1\|_\infty + \sqrt{n}\|\text{Rem}_2\|_\infty = o_p\left(\sqrt{Q_{kk}}\right). \]

First, using \((\tilde{\Sigma}^z)^{-1} - (\Sigma^z)^{-1} = (\tilde{\Sigma}^z - \Sigma^z)(\Sigma^z)^{-1}\) bounds \(\sqrt{n}\|\text{Rem}_2\|_\infty\) from above by

\[ \sqrt{n}\|\tilde{\Sigma}^z - \Sigma^z\|_{op} = O_p\left(1\right). \]

From display (75) and Lemmas 17 and 18 in Section B.7, applied with \(\rho = 0, u = e_k\) and \(v = \beta\), and by replacing \(\tilde{\Sigma}\) with \(\Sigma\), we obtain

\[ \sqrt{n}\|\tilde{\Sigma}^w\beta + \hat{X}^T\varepsilon - \frac{1}{n}Z^T\tilde{\Sigma}\beta\|_\infty = O_p\left(1\right). \]

On the other hand, write \(\Omega = (\tilde{\Sigma}^z)^{-1}\). For any \(k \in [K]\), we have

\[ \|\Omega(\tilde{\Sigma}^z - \Sigma^z)\Omega_k\|_1 \leq \sqrt{K}\|\Omega(\tilde{\Sigma}^z - \Sigma^z)\Omega_k\|_1 \leq \sqrt{K}\|\Omega\|_{op}\|\tilde{\Sigma}^z - \Sigma^z\|_{op}\|\Omega_k\|. \]

Note that \(\|\Omega\|_{op} = C^{-1}_{\min}\) and \(\|\Omega_k\| \leq \|\Omega\|_{op} = \lambda^{-1}_{\min}(\tilde{\Sigma}^z) = O_p(C^{-1}_{\min})\) from (72). Invoking Lemma 21 in Section B.7 and Assumption 3 concludes

\[ \|\Omega(\tilde{\Sigma}^z - \Sigma^z)\Omega_k\|_1 = O_p\left(C_{\max}C_{\min}^{-2}K\sqrt{\log(p \vee n)/n}\right). \]

Hence, we have

\[ \sqrt{n}\|\text{Rem}_2\|_\infty = O_p\left(C_{\max}^{-1}\sqrt{K}\log(p \vee n)/n\right). \]

For the term \(\text{Rem}_1\), invoking Lemma 22 in Section B.7 yields

\[ \sqrt{n}\|\text{Rem}_1\|_\infty = O_p\left(C_{\min}^{-1}\left(1 \vee \sqrt{\frac{\|\beta\|}{m}}\right)\sqrt{K}\log(p \vee n)/n\right). \]

Finally, using \(\Omega_{kk} \geq \lambda_{\min}(\Omega) = C^{-1}_{\max}\) and \(\kappa(\Sigma^z) = C_{\max}/C_{\min}\), we have

\[ \sqrt{n}\|\text{Rem}_2\|_\infty/\sqrt{Q_{kk}} = O_p\left(\kappa^2(\Sigma^z)K\log(p \vee n)/\sqrt{n}\right) = o_p(1) \]

provided that \(K\log(p \vee n) = o(\sqrt{n})\), and

\[ \sqrt{n}\|\text{Rem}_1\|_\infty/\sqrt{Q_{kk}} = O_p\left(\kappa(\Sigma^z)\rho K\sqrt{m}\log(p \vee n)\right) = o_p(1), \]

by Assumption 3'. This completes the proof.

**B.7 Lemmas used in the proof of Theorems 4, 5 and 16**

**Lemma 17.** Let \(\tilde{Z}\) and \(\tilde{W}\) be defined in (67) for any given output \(\hat{f}\). Let \(u, v \in \mathbb{R}^K\) be any fixed vectors. Under the conditions of Theorem 2, with probability \(1 - (p \vee n)^{-1/2}\), one has

\[ \frac{1}{n}\|\tilde{Z}^T\tilde{W}v\|_\infty \lesssim \left(\frac{\|v\|}{\sqrt{m}} \vee \rho\|v\|_1\right) \sqrt{\frac{\log(p \vee n)}{n}}. \]  
\[ \frac{1}{n}\|u^T\tilde{Z}^T\tilde{W}v\| \lesssim \left(\frac{\|v\|}{\sqrt{m}} \vee \rho\|u\|_1\|v\|_1\right) \sqrt{\frac{\log(p \vee n)}{n}}. \]
Proof. This lemma is proved in Section B.14.1.

**Lemma 18.** Let \( \widetilde{W} \) and \( D = \text{diag}(d_1, \ldots, d_K) \) be defined in (67) and (74), respectively. Under the conditions of Theorem 2, for any fixed vectors \( u, v \in \mathbb{R}^K \), with probability \( 1 - (p \vee n)^{-c} \),

\[
\left| u^T \left( \frac{1}{n} \widetilde{W}^T \widetilde{W} - D \right) v \right| \lesssim \left( \frac{\|u\|}{\sqrt{m - 1}} + \rho \|u\|_1 \right) \left( \frac{\|v\|}{\sqrt{m}} + \rho \|v\|_1 \right) \sqrt{\frac{\log(p \vee n)}{n}}.
\]

**Proof.** This lemma is proved in Section B.14.2.

**Lemma 19.** Let \( \widetilde{X} \) and \( \widetilde{Z} \) be defined as (67). Under the conditions of Theorem 2, one has

\[
\mathbb{P} \left\{ \frac{1}{n} \| \widetilde{X}^T (Z - \widetilde{Z}) \beta \|_\infty \lesssim \frac{\rho \| \beta \|_1}{\sqrt{\log(p \vee n)}} \frac{\log(p \vee n)}{n} \right\} \geq 1 - (p \vee n)^{-c}.
\]

**Proof.** This lemma is proved in Section B.14.3.

**Lemma 20.** Under the conditions of Theorem 2, for any fixed vector \( v, u \in \mathbb{R}^K \), we have

\[
\left| u^T (\hat{\Sigma}^z - \Sigma^z) v \right| \lesssim \left( \sqrt{u^T \Sigma^z u} \sqrt{v^T \Sigma^z v} + \rho \|u\|_1 \|v\|_1 \right) \sqrt{\frac{\log(p \vee n)}{n}}
\]

with probability greater than \( 1 - (p \vee n)^{-c} \) for some constant \( c > 0 \).

**Proof.** This lemma is proved in Section B.14.4.

**Lemma 21.** Under the conditions of Theorem 2, we have

\[
\mathbb{P} \left\{ \| \hat{\Sigma}^z - \Sigma^z \|_{op} \lesssim \left( C_{\max} \vee \rho \sqrt{K} \right) \sqrt{\frac{K \log(p \vee n)}{n}} \right\} \geq 1 - (p \vee n)^{-c},
\]

\[
\mathbb{P} \left\{ \left\| \frac{1}{n} Z^T Z - \Sigma^z \right\|_{op} \lesssim C_{\max} \sqrt{\frac{K \log(p \vee n)}{n}} \right\} \geq 1 - O(p^{-K}).
\]

**Proof.** This lemma is proved in Section B.14.5.

**Lemma 22.** Under conditions of Theorem 5, let \( \bar{X} \) and \( \bar{\Sigma}^z \) be defined in the proof of Theorem 5. Then, with probability greater than \( 1 - (p \vee n)^{-c} \) for some constant \( c > 0 \),

\[
\left\| (\hat{\Sigma}^z)^{-1} \left[ \frac{1}{n} (\bar{X} - \bar{X})^T Y - (\hat{\Sigma}^z - \Sigma^z) \beta \right] \right\|_\infty \lesssim C_{\min}^{-1} \rho (1 \vee \sqrt{K} \| \beta \|) \sqrt{\frac{K \log(p \vee n)}{n}}.
\]

**Proof.** This lemma is proved in Section B.14.6.
B.8 Proof of Proposition 6: Fisher information bound

We start with the following joint distribution of \((X,Y)\) from model (1),

\[
\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_{p+1} \left( 0, \begin{bmatrix} \Sigma & A \Sigma^z \beta \\ \beta^T \Sigma^z A^T & \beta^T \Sigma^z \beta + \sigma^2 \end{bmatrix} \right).
\]

This implies \(Y|X \sim N(\mu_{y|x}, \sigma^2_{y|x})\) where \(\mu_{y|x} = X^T \Sigma^{-1} A \Sigma^z \beta\) and

\[
\sigma^2_{y|x} = \sigma^2 + \beta^T \Sigma^z \beta - \beta^T \Sigma^z A \Sigma^{-1} A \Sigma^z \beta = \sigma^2 + \beta^T G^{-1} \beta
\]

from Fact 1. Since the probability density of \(X\) does not depend on \(\beta\), we have \(I_{X,Y}(\beta) = I_{Y|X}(\beta)\) for the Fisher information. Using the log-likelihood of \(Y|X\),

\[
2 \ell_{Y|X}(\beta) = -\log \sigma^2_{y|x} - \sigma^2_{y|x} (Y - \mu_{y|x})^2 + \text{constant},
\]

a basic calculation shows that

\[
I_{Y|X}(\beta) = \sigma^2_{y|x} \left[ \Sigma^z A^T \Sigma^{-1} A \Sigma^z + 2 \sigma^2_{y|x} G^{-1} \beta \beta^T G^{-1} \right].
\]

Let \(M = \Sigma^z A^T \Sigma^{-1} A \Sigma^z\) and \(v = G^{-1} \beta\). An application of Woodbury matrix identity gives

\[
\left[ M + 2 \sigma^2_{y|x} G^{-1} \beta \beta^T G^{-1} \right]^{-1} = M^{-1} - \frac{2M^{-1} \sigma^2_{y|x} G^{-1} \beta \beta^T G^{-1}}{\sigma^2_{y|x} + 2v^T M^{-1} v}
\]

\[
= M^{-1} \left[ I_K - \frac{2M^{-1/2} vv^T M^{-1/2}}{\sigma^2_{y|x} + 2v^T M^{-1} v} \right].
\]

For any \(k \in [K]\), this implies

\[
\left[ I_{X,Y}^{1}(\beta) \right]_{kk} = \sigma^2_{y|x} [M^{-1}]_{kk} \left[ 1 - \frac{2e_k^T M^{-1/2} vv^T M^{-1/2} e_k}{\sigma^2_{y|x} + 2v^T M^{-1} v} \right]
\]

\[
\geq \sigma^2_{y|x} [M^{-1}]_{kk} \frac{\sigma^2 + \beta^T G^{-1} \beta}{\sigma^2 + \beta^T G^{-1} \beta + 2v^T M^{-1} v}
\]

where in the second line we use

\[
e_k^T M^{-1/2} vv^T M^{-1/2} e_k \leq \|M^{-1/2} vv^T M^{-1/2}\|_{op} = v^T M^{-1} v
\]

and plug in \(\sigma^2_{y|x} = \sigma^2 + \beta^T G^{-1} \beta\). Observe that Fact 1 yields

\[
M^{-1} = (\Sigma^z - G^{-1})^{-1} = (\Omega^{-1}(G - \Omega)G^{-1})^{-1} = \Omega + \Omega(A^T \Omega^w A)^{-1} \Omega.
\]

Then to prove the first result, it suffices to upper bound \(v^T M^{-1} v\). Since

\[
v^T M^{-1} v = \beta^T G^{-1/2} \left( G^{-1/2} M^{-1} G^{-1/2} \right) G^{-1/2} \beta
\]

\[
\leq \beta^T G^{-1} \beta \left[ \lambda_{\min}(G^{1/2} MG^{1/2}) \right]^{-1}.
\]

47
and, by using $M = \Sigma - G^{-1}$ from Fact 1 and $G = \Omega + A^T\Omega^w A$,

$$
\lambda_{\min}(G^{1/2}MG^{1/2}) = \lambda_{\min}(G^{1/2}\Sigma G^{1/2}) - 1 = \lambda_{\min}\left((\Sigma^{1/2})A^T\Omega^w A(\Sigma^{1/2})\right),
$$

the first result follows by observing that

$$
\lambda_{\min}\left((\Sigma^{1/2})A^T\Omega^w A(\Sigma^{1/2})\right) \geq \frac{\lambda_K(A\Sigma^z A^T)}{||\Sigma^w||_{op}}.
$$

When $\lambda_1(A^T A) = c_1 m$, we first notice that $\lambda_K(A\Sigma^z A^T) \geq C_{\min}\lambda_K(A^T A) \geq C_{\min}\lambda_K(A_I^T A_I) = mC_{\min}$. Then the result follows by

$$
||G||_{op} \leq ||\Omega||_{op} + ||A^T\Omega^w A||_{op} = C_{\min} - 1 + c_1 m/\tau^2 = (c_0 + c_1)m/\tau^2
$$

and $\beta^T G^{-1} \beta \geq ||\beta||^2 ||G||_{op}^{-1}$.

\[\Box\]

**B.9 Proof of Proposition 7: consistent estimation of $Q_{kk}$**

We only need to show the consistency of $\widehat{Q}_{kk}$ since the rest of the proof follows from Theorem 5 and an application of the Slutsky’s theorem.

By Taylor expansion, we have

$$
\left|\frac{\widehat{Q}_{kk}^{1/2}/Q_{kk}^{1/2} - 1}{Q_{kk}^{-1}}\right| = Q_{kk}^{-1} \left|\widehat{Q}_{kk} - Q_{kk}\right| (1 + o_p(1)),
$$

provided that $Q_{kk}^{-1} \widehat{Q}_{kk} - Q_{kk} = o_p(1)$. Thus, to show $\left|\frac{\widehat{Q}_{kk}^{1/2}/Q_{kk}^{1/2} - 1}{Q_{kk}^{-1}}\right| = o_p(1)$ it suffices to show $Q_{kk}^{-1} \widehat{Q}_{kk} - Q_{kk} = o_p(1)$. To this end, begin by recalling that

$$
Q_{kk} = \left(\sigma^2 + \frac{\tau^2||\beta||^2}{m}\right)\left(\Omega_{kk} + \frac{\tau^2\|\Omega_k^*\|^2}{m}\right) + \frac{\tau^4\sum a \Omega_{ka}^2 \beta_a^2}{m^2(m-1)}.
$$

Note that the last term is obviously smaller than $\Delta_1\Delta_2$ since it is a product of two term, each of which is smaller than $\Delta_1$ and $\Delta_2$, respectively. It suffices to bound $|\Delta_1\Delta_2 - \Delta_1\Delta_2|$ with

$$
\Delta_1 := \hat{\sigma}^2 + \frac{\tau^2||\beta||^2}{m}, \quad \Delta_2 := \hat{\Omega}_{kk} + \frac{\tau^2\|\hat{\Omega}_k^*\|^2}{m}
$$

where we can take $\hat{m} = [\hat{I}_k]$ and $\hat{\tau}^2 = \hat{\tau}_i^2$ for any arbitrary $1 \leq k \leq K$ and $1 \leq i \leq p$. We thus have

$$
\left|\frac{\widehat{Q}_{kk} - Q_{kk}}{Q_{kk}}\right| \leq \frac{|\Delta_1 - \Delta_1|}{\Delta_1} + \frac{|\Delta_2 - \Delta_2|}{\Delta_2} + \frac{|\Delta_1 - \Delta_1||\Delta_2 - \Delta_2|}{\Delta_1\Delta_2}, \quad (97)
$$

On the one hand, from the definition of $\rho$ in (17),

$$
|\Delta_1 - \Delta_1|
\leq |\sigma^2 - \hat{\sigma}^2| + \frac{\tau^2||\beta||^2}{m} \left|\frac{m - \hat{m}}{m}\right| + \frac{||\beta||^2}{m} |\tau^2 - \hat{\tau}^2| + \frac{\tau^2}{m} (||\beta|| + ||\beta||)||\hat{\beta} - \beta||
\leq |\sigma^2 - \hat{\sigma}^2| + \frac{\tau^2||\beta||^2}{m} \rho + \frac{||\beta||^2}{m} |\tau^2 - \hat{\tau}^2| + \frac{\tau^2}{m} (||\beta|| + ||\beta||)||\hat{\beta} - \beta||.
$$

48
Note that (23), $K \log(p \lor n) = o(n)$ and the rate of $\max_{i \in [p]} |\hat{\tau}_i^2 - \tau_i^2|$ in Lemma 23 in Section B.10 guarantee that
\begin{equation}
\|\hat{\beta}\| = O(1 \lor \|\beta\|), \quad \hat{\tau}^2 = O(\tau^2)
\end{equation}
with probability greater than $1 - (p \lor n)^{-c}$. Invoking Lemmas 23 and 24 in Section B.10 together with the rate of $\|\hat{\beta} - \beta\|$ from (23) further gives
\[
\mathbb{P}\left\{|\hat{\Delta}_1 - \Delta_1| \lesssim \left(1 \lor \frac{\|\beta\|^2}{m}\right) \left(\rho + C_{\min}^{-1} \sqrt{\frac{K \log(p \lor n)}{n}}\right)\right\} \geq 1 - (p \lor n)^{-c}.
\]
With the same probability, under $K \log(p \lor n) = o(n)$ and Assumption 3, plugging $\Delta_1$ into the above display yields
\[
\Delta_1^{-1}|\hat{\Delta}_1 - \Delta_1| \lesssim \rho + C_{\min}^{-1} \sqrt{K \log(p \lor n)/n} = o(1).
\]
On the other hand,
\[
|\hat{\Delta}_2 - \Delta_2| \leq |\hat{\Omega}_{kk} - \Omega_{kk}| + \hat{\tau}^2 \frac{\|\hat{\Omega}_k\|^2}{m} \rho + \frac{\|\Omega_k\|^2}{m} |\hat{\tau}^2 - \tau^2| + \frac{\hat{\tau}^2}{m} (\|\Omega_k\| + \|\hat{\Omega}_k\|) \|\hat{\Omega}_k - \Omega_k\|.
\]
Recall that $\hat{\Omega} - \Omega = \Omega(\Sigma^2 - \hat{\Sigma}^2)\hat{\Omega}$. We thus have
\[
\|\hat{\Omega}_k - \Omega_k\| \lesssim \|\Omega\|_{op} \|\Sigma^2 - \hat{\Sigma}^2\| \Omega_k \| \leq \lambda_{\min}^{-1}(\hat{\Sigma}^2) \sqrt{K} \|\Sigma^2 - \Sigma^2\| \Omega_k \|_{\infty}.
\]
From Lemma 20 in Section B.7 with $\rho$ satisfying Assumption 3 and by choosing $u = e_a$ and $v = \Omega_k$ for any $a, k \in [K]$, we obtain
\[
\mathbb{P}\left\{|e_a^T (\Sigma^2 - \hat{\Sigma}^2) \Omega_k | \lesssim \sqrt{\Omega_{kk} \Sigma_{aa}^z} \sqrt{\log(p \lor n)/n}\right\} \geq 1 - (p \lor n)^{-c}.
\]
Since $\Omega_{kk} \leq \|\Omega\|_{op} \leq C_{\min}^{-1}$, $\Sigma_{aa}^z = O(1)$ from Assumption 2 and $\lambda_{\min}^{-1}(\hat{\Sigma}^2) \lesssim C_{\min}^{-1}$ with the same probability, taking the union bound over $a \in [K]$ and invoking (72) give
\[
\mathbb{P}\left\{\|\hat{\Omega}_k - \Omega_k\| \lesssim C_{\min}^{-3/2} \sqrt{K \log(p \lor n)/n}\right\} \geq 1 - (p \lor n)^{-c},
\]
which also upper bounds $|\hat{\Omega}_{kk} - \Omega_{kk}|$. By noting that $\|\Omega_k\| \geq \lambda_{\min} (\Omega) = C_{\max}^{-1}$, under $K \log(p \lor n) = o(n)$, from the previous display we also find
\[
\|\hat{\Omega}_k\| \leq \|\Omega_k\| + \|\hat{\Omega}_k - \Omega_k\| = O_p(\|\Omega_k\|).
\]
Therefore, invoking Lemma 23 in Section B.10, using $\hat{\tau}^2 = O_p(\tau^2)$ from (98) and collecting (100) – (101) give
\[
\Delta_2^{-1}|\hat{\Delta}_2 - \Delta_2| \lesssim \rho + C_{\min}^{-3/2} \left(\sqrt{\frac{\|\Omega\|_{\infty}}{m}} + \sqrt{\frac{\|\Omega\|_{\infty, 2}}{m}}\right) \sqrt{\frac{K \log(p \lor n)}{n}} = o(1)
\]
provided that $\|\Omega\|_{\infty} \leq \|\Omega\|_{\infty, 2} \leq C_{\min}^{-1}$. Since the third term in (97) is the product of the first and second terms, which we have shown converging to zero, this concludes $Q_{kk}^{-1}|\hat{Q}_{kk} - Q_{kk}| = o_p(1)$ and completes the proof. \qed

49
B.10 Lemmas used in the proof of Proposition 7

Lemma 23. Under the same conditions of Theorem 2, assume \( \min_i \tau_i^2 \geq C_W > 0 \) for some constant \( C_W \). Let \( \hat{\tau}_i^2 \) be defined in (27). With probability greater than \( 1 - (p \lor n)^{-c} \) for some constant \( c > 0 \), we have

\[
\max_{i \in [p]} |\hat{\tau}_i^2 - \tau_i^2| \lesssim \sqrt{\frac{\log(p \lor n)}{n}}, \quad \max_{i \in [p]} |\hat{\tau}_i - \tau_i| \lesssim C_W^{-2} \sqrt{\frac{\log(p \lor n)}{n}}.
\]

Proof. This lemma is proved in Section B.15.1. \( \square \)

Lemma 24. Let \( \hat{\sigma}^2 \) be defined as in (28). Under the conditions of Theorem 4, one has

\[
P \left\{ |\hat{\sigma}^2 - \sigma^2| \leq C^{-1}_{\min} \left( 1 \lor \frac{\|\beta\|^2}{m} \right) \sqrt{\frac{K \log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c},
\]

for some constant \( c > 0 \).

Proof. This lemma is proved in Section B.15.2 \( \square \)

Lemma 25. Under the same conditions of Theorem 2, let \( \hat{\Sigma}^z \) be constructed as (53) and \( \hat{A} \) as in display (56) together with (54) - (55). Let \( s_j = \|A_j\|_0 \) for any \( j \in [p] \) and \( C_W = \max_{i \in [p]} \Sigma_{ii}^w \). With probability greater than \( 1 - (p \lor n)^{-c} \) for some constant \( c > 0 \),

\[
\|\hat{A}_j - A_j\| \lesssim C^{-1}_{\min} \sqrt{\frac{C_W s_j \log(p \lor n)}{n}}, \quad \|\hat{\Sigma}^z(\hat{A}_j - A_j)\|_\infty \lesssim \sqrt{\frac{C_W \log(p \lor n)}{n}}
\]

hold for all \( 1 \leq j \leq p \).

Proof. Write \( s := s_j \). The rate of \( \hat{A}_j - A_j \) follows immediately from (Bing et al., 2017, Theorem 5 display (36)) by observing that

\[
\kappa_2(\Sigma^z, s) := \inf_{|S| \leq s} \inf_{\|v\|=1} \|\Sigma^z v\|_\infty \geq \inf_{|S| \leq s} \inf_{\|v\|=1} \frac{\|\Sigma^z v\|_\infty}{\|S v_S\|_\infty} \geq \inf_{|S| \leq s} \inf_{\|v\|=1} \frac{\|\Sigma^z v_S\|_\infty}{\|S v_S\|_\infty} \geq \inf_{|S| \leq s} \inf_{\|v\|=1} \frac{\|\Sigma^z v_S\|_\infty}{\sqrt{s}} \geq C_{\min}/\sqrt{s}
\]

with \( C_S := \{ v \in \mathbb{R}^K : \|v_S\|_1 \leq \|v_S\|_1 \} \) and \( S \subseteq [K] \) with \( |S| \leq s \). Letting \( \hat{H}^j = (\hat{A}_j^T \hat{A}_j)^{-1} \hat{A}_j^T \hat{\Sigma}_j^w \), the proof of \( \|\hat{\Sigma}^z(\hat{A}_j - A_j)\|_\infty \) follows from

\[
\|\hat{\Sigma}^z(\hat{A}_j - A_j)\|_\infty \leq \|\hat{\Sigma}^z \hat{A}_j - \hat{H}^j\|_\infty + \|\hat{\Sigma}^z A_j - \hat{H}^j\|_\infty = O \left( \sqrt{C_W \log(p \lor n) / n} \right)
\]

by using the feasibility of both \( \hat{A}_j \) and \( A_j \). \( \square \)
\textbf{B.11 Proof of Theorem 8: the in-sample and new-data prediction risks}

We first need some notations. Define the quantities
\begin{equation}
R_A := C_{\min}^{-1} \sqrt{C_W \|A_I\|_O \log (p \lor n)/n}, \quad R_M := R_A \Lambda_{\min}^{-1/2}.
\end{equation}
\begin{equation}
C_{w,j} = C_W [1 \lor (|J| \log (p \lor n))/n],
\end{equation}
by recalling that \( \Lambda_{\min} := \lambda_\nu(A^T A) \) and \( C_W = \max_{1 \leq i \leq \nu} \Sigma_{ii}^w \). We further write \( M_I = A_I^T A_I \) and \( M = A^T A \) in the sequel for notational simplicity.

We then give three crucial lemmas which are used repeatedly in this section. Their proofs are deferred to Section B.16.1, B.16.2 and B.16.3.

\textbf{Lemma 26.} Let Assumption 1 holds. We have
\begin{enumerate}[(i)]
\item \( \lambda_\nu(M_I) = m, \|A_I M_I^{-1}\|_{op} = \sqrt{m}; \)
\item \( \|AM^{-1}\|_{op} = \Lambda_{\min}^{-1/2}, \|AM^{-1/2}\|_{op} = 1. \)
\end{enumerate}

\textbf{Lemma 27.} Let Assumption 4 and \( K \log (p \lor n) = O(n) \) hold. With probability greater than \( 1 - (p \lor n)^{-c} \) for some constant \( c > 0 \), we have
\begin{enumerate}[(i)]
\item \( nC_{\min} \lesssim \lambda_\nu(Z^T Z) \lesssim \lambda_{\max}(Z^T Z) \lesssim nC_{\max}; \)
\item \( \lambda_{\min}(M^{-1} A_1 X^T XAM^{-1}) \gtrsim nC_{\min}; \)
\item \( \|W_I\|_{op} \lesssim nC_{w,j}; \)
\item \( \|WAM^{-1/2}\|_{op} \lesssim nC_W. \)
\end{enumerate}

\textbf{Lemma 28.} Let \( \hat{M} = \hat{A}^T \hat{A} \). Under the conditions of Theorem 8 and \( s_J \log (p \lor n) = o(n\Lambda_{\min}) \), on the event \( \mathcal{E}_f \) defined in (105), we have
\begin{enumerate}[(i)]
\item \( \|\hat{M}^{-1}\|_{op} \lesssim \Lambda_{\min}^{-1}, \|\hat{A}M^{-1}\|_{op} \lesssim \Lambda_{\min}^{-1/2}; \)
\item \( \|WAM^{-1} - WAM^{-1}\|_{op} \lesssim nR_M^2; \)
\item \( \|X \hat{A} \hat{M}^{-1} - XAM^{-1}\|_{op} \lesssim nC_{\max}R_M^2. \)
\end{enumerate}

We first prove the result in (35) and then, under \( s_J \log (p \lor n) = o(n\Lambda_{\min}) \), prove (36) for both \( \mathbb{E}\|\hat{Y} - Z\hat{\beta}\|^2 \) and \( \mathbb{E}[(\hat{Y}_* - Z^T_* \hat{\beta})^2 \land T] \).

\textbf{B.11.1 Proof of in sample prediction risk in (35)}

We write \( \hat{Z} := X \hat{A} \) and \( \hat{\beta}^{LS} = (\hat{Z}^T \hat{Z})^+ \hat{Z}^T Y \) such that \( \hat{Y} = \hat{Z} \hat{\beta}^{LS} \), where \( M^+ \) denotes the Moore-Penrose inverse of any matrix \( M \). Let \( \hat{P} = \hat{Z}(\hat{Z}^T \hat{Z})^+ \hat{Z}^T \) be the orthogonal projection onto the column space of \( \hat{Z} \) and \( \hat{P}^\perp = I_n - \hat{P} \). Using \( Y = Z\beta + \varepsilon \) gives
\begin{equation}
\mathbb{E}\|Z\beta - \hat{Z} \hat{\beta}^{LS}\|^2 = \mathbb{E}\|Z\beta - \hat{P} Y\|^2 = \mathbb{E}\|\hat{P} \perp Z\beta\|^2 + \mathbb{E}\|\hat{P}\varepsilon\|^2.
\end{equation}
Recalling that (102) and (103), define the event
\[ \mathcal{E}_f = \left\{ \tilde{K} = K, \ I_k \subseteq \tilde{I}_k \subseteq I_k \cup J_k, \forall k \in [K], \ \|\tilde{A} - A\|_{op} \lesssim R_A \right\} \]
(105)
\[ nC_{min} \lesssim \lambda_{min}(Z^T Z) \leq \lambda_{max}(Z^T Z) \lesssim nC_{max}, \ \lambda_{min}(\hat{\Sigma}) \gtrsim C_{min}, \]
\[ \|WM^{-1/2}\|_{op}^2 \lesssim nC_W, \ \lambda_{max}(W_{J}^T W_{J}) \lesssim nC_{wJ}. \]
We have \( \mathbb{P}(\mathcal{E}_f) \geq 1 - (p \lor n)^{-c} \) for some constant \( c > 0 \), from Theorem 2, Lemma 21 in Section B.7, Lemma 25 and Lemma 27.

We first study \( \mathbb{E}\|\hat{P} Z\beta\|^2 \) on the event \( \mathcal{E}_f \). Recalling that \( M = A^T A \) and using \( \hat{Z} = X\hat{A} = ZA^T\hat{A} + W\hat{A} \) give
\[
\|\hat{P} Z\beta\|^2 = \|\hat{P} (Z - \hat{Z}M^{-1})\beta\|^2 \\
\leq 2\|Z - ZA^T\hat{A}M^{-1}\|\beta\|^2 + 2\|W\hat{A}M^{-1}\|\beta\|^2 \\
\leq 2\|Z\|^2_{op}\|\hat{A}\|_{op}^2\|\hat{A} - A\|^2_{op}\|\beta\|^2 + 4\|WAM^{-1}\|\beta\|^2 + 4\|W(\hat{A} - A)M^{-1}\|\beta\|^2 \\
\leq 2\|Z\|^2_{op}\|\hat{A}\|_{op}^2\|\beta\|^2 + 4\|WAM^{-1}\|\beta\|^2 + 4\|WAM^{-1}\|\beta\|^2 \\
+ 4\|Wj\|_{op}^2\|\hat{A} - A\|^2_{op}\|\beta\|^2 \quad (\text{by } \hat{A}_I = A_I) \\
\lesssim nC_{max}\|\beta\|^2 R_M^2 \kappa_A + 4\|WAM^{-1}\|\beta\|^2 + nC_{wJ}\|\beta\|^2 R_M^2 \Lambda_{min}^{-1}
\]
by invoking \( \mathcal{E}_f \) in the last line, where \( \kappa_A = \lambda_1(A^T A) / \lambda_K(A^T A) = \|A\|_{op}^2 \|M^{-1}\|_{op} \). By taking the expectation, we thus have
\[
\mathbb{E}\left[ \|\hat{P} Z\beta\|^2_{1_{\mathcal{E}_f}} \right] \lesssim n\|\beta\|^2 R_M^2 (C_{max}\kappa_A + C_{wJ} \Lambda_{min}^{-1}) + n\beta^2 M^{-1} A^T \Sigma^w A M^{-1}\beta,
\]
which, by \( \|\Sigma^w\|_{\infty} \leq C_W \), further yields
\[
\mathbb{E}\left[ \|Z\beta - \hat{Z} \beta^{LS}\|^2_{1_{\mathcal{E}_f}} \right] \leq \mathbb{E}\left[ \|\hat{P}\| E_1\|X\| \right] \leq \sigma^2 \mathbb{E}\left[ \|\tilde{K} \|_{1_{\mathcal{E}_f}} \right] \leq K \sigma^2.
\]
Noting that \( \mathcal{E}_f \) and \( \hat{P} \) only depend on \( X \), the first term can be analyzed as
\[
\mathbb{E}\left[ \|\hat{P}\| E_1\|X\| \right] = \mathbb{E}\left[ \text{trace}\left( \hat{P} E_1\|X\| \right) \right] \leq \sigma^2 \mathbb{E}\left[ \|\tilde{K} \|_{1_{\mathcal{E}_f}} \right] \leq K \sigma^2.
\]
We finally obtain
\[
\left( \frac{1}{n} \mathbb{E}\left[ \|Z\beta - \hat{Z} \beta^{LS}\|^2_{1_{\mathcal{E}_f}} \right] \right) \lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2}{\Lambda_{min}} C_W + \|\beta\|^2 R_M^2 (C_{max}\kappa_A + C_{wJ} \Lambda_{min}^{-1}).
\]
To conclude the proof, from (104), we have
\[
\mathbb{E}\left[ \|Z\beta - \hat{Z} \beta^{LS}\|^2_{1_{\mathcal{E}_f}} \right] \leq \mathbb{E}\left[ \|\hat{P} Z\beta\|^2_{1_{\mathcal{E}_f}} \right] + \mathbb{E}\left[ \|\hat{P}\| E_1\|X\| \right] \\
\leq \mathbb{E}\left[ \|Z\beta\|^2_{1_{\mathcal{E}_f}} \right] + \mathbb{E}\left[ \|\epsilon\|^2_{1_{\mathcal{E}_f}} \right] \\
\leq (\mathbb{E}\left[ \|Z\beta\|^3 \right] )^{2/3} (\mathbb{P}(\mathcal{E}_f^c)^{1/3} + n\sigma^2 \mathbb{P}(\mathcal{E}_f^c).}
The last inequality uses the independence between \(1_{E_f}\) and \(\varepsilon\) and the Hölder’s inequality. Note that

\[
\|Z\beta\| = \left(\sum_{i=1}^{n} U_i^2\right)^{1/2} = \left(\sum_{i=1}^{n} (z_i, \beta)^2\right)^{1/2}.
\]

The fact that, \(U_1, \ldots, U_n\) are independent and \((\|\beta\|\gamma \sqrt{C_{max}})-\)subgaussian, implies that that \(\|Z\beta\| = \left(\sum_{i=1}^{n} U_i^2\right)^{1/2}\) is \((nc')\gamma \sqrt{C_{max}})-\)subgaussian, for a constant \(c' > 0\) (see Lemmas 5.9 and 5.14 in Vershynin (2012)). We thus obtain that

\[
E\|Z\beta\|^3 \lesssim \gamma^3 C_{max}^3 \|\beta\|^3 n^{3/2} = O(K^{3/2} n^{3/2}), \quad (by \ |\beta\| = O(\sqrt{K}))
\]

which, together with \(P(E_f^c) = (p \lor n)^{-c}\), concludes

\[
\frac{1}{n} E \left[\|Z\beta - \hat{Z}\beta LS\|^2 1_{E_f}\right] \lesssim C_{max} K (p \lor n)^{-c/3} + \sigma^2 (p \lor n)^{-c}.
\]

Taking \(c\) large enough completes the proof for \(\hat{Y}\).

\[\Box\]

**B.11.2 Proof of in sample prediction risk in (36)**

We use the same arguments as before under the additional assumption, \(s_f \log(p \lor n) = o(n \Lambda_{min})\), which further implies

\[
R_M = C_{W}^{1/2} \sqrt{s_f \log(p \lor n) \over n \Lambda_{min}} = o(1).
\]

(106)

To prove (36), by the previous arguments of the proof of (35) and writing \(\hat{M} = \hat{A}^T \hat{A}\), one can derive

\[
\begin{align*}
\|\hat{P} Z\beta\|^2 &= \|\hat{P}^\perp (Z - \hat{Z} \hat{M}^{-1}) \beta\|^2 \\
&\leq 2\|(Z - Z A^T \hat{A} \hat{M}^{-1}) \beta\|^2 + 2\|W \hat{A} \hat{M}^{-1} \beta\|^2 \\
&\leq 2\|Z (A - \hat{A})^T \hat{A} \hat{M}^{-1} \beta\|^2 + 4\|W A M^{-1} \beta\|^2 + 4\|W (\hat{A} \hat{M}^{-1} - A M^{-1}) \beta\|^2 \\
&\leq 2\|Z T Z \|_{op} \|\hat{A} - A\|_{op}^2 \|\hat{A} M^{-1} \|_{op}^2 \|\beta\|^2 + 4\|W A M^{-1} \beta\|^2 + nR_M^2 \|\beta\|^2 \\
&\lesssim nC_{max} \|\beta\|^2 2R_M^2 + 4\|W A M^{-1} \beta\|^2
\end{align*}
\]

where we use part (ii) of Lemma 28 in the fourth line and invoke \(E_f\) with part (i) of Lemma 28 in the last line. This further leads to

\[
E \left[\|Z\beta - \hat{Z}\beta LS\|^2 1_{E_f}\right] \leq E \left[\|\hat{P} \varepsilon\|^2 1_{E_f}\right] + nC_{max} \|\beta\|^2 2R_M^2 + n\|\beta\|^2 \Lambda_{min}^{-1} C_W
\]

\[
\leq K \sigma^2 + nC_{max} \|\beta\|^2 2R_M^2 + n\|\beta\|^2 \Lambda_{min}^{-1} C_W.
\]

The same arguments can be used to upper bound \(E[\|Z\beta - \hat{Z}\beta LS\|^2 1_{E_f}]\) from which we complete the proof for (36).

\[\Box\]
B.11.3 Proof of (36) for $\mathbb{E}[Z_\ast \beta - \hat{Y}_\ast]^2$

Recall that $M = A^T A$ and $\hat{Y}_\ast = \hat{\theta}^T X_\ast$ with $\hat{\theta}$ defined in (32). Define the event

$$\mathcal{E}_p = \mathcal{E}_f \cap \left\{ \lambda_{\min}(M^{-1}A^T X^T X A M^{-1}) \gtrsim n C_{\min}, \quad \|W A M^{-1}\|_{op} \lesssim n \frac{C_W}{\Lambda_{\min}} \right\}$$

with the event $\mathcal{E}_f$ defined in (105). From (ii) and (iv) of Lemma 27, we know $\mathbb{P}(\mathcal{E}_p) = (p \lor n)^{-c}$ for some constant $c > 0$, which yields

$$\mathbb{E}\left[ \left( \|Z_\ast \beta - \hat{Y}_\ast\| \right) \cdot 1_{\mathcal{E}_p} \right] \leq T \cdot \mathbb{P}(\mathcal{E}_p) \lesssim (p \lor n)^{-c}. \quad (107)$$

Thus, it suffices to upper bound $\mathbb{E}[\|Z_\ast \beta - \hat{Y}_\ast\|^2 1_{\mathcal{E}_p}]$. From $X_\ast = AZ_\ast + W_\ast$ and the independence of $Z_\ast$ and $W_\ast$, by writing $\Delta = \beta - A^T \hat{\theta}$, we have

$$\mathbb{E}[\|Z_\ast \beta - \hat{Y}_\ast\|^2 1_{\mathcal{E}_p}] = \mathbb{E}\left[ (\beta - A^T \hat{\theta})^T Z_\ast \cdot 1_{\mathcal{E}_p} \right]^2 + \mathbb{E}\left[ \hat{\theta}^T W_\ast \cdot 1_{\mathcal{E}_p} \right]^2$$

$$= \mathbb{E}\left[ \Delta^T \Sigma Z \cdot 1_{\mathcal{E}_p} \right] + \mathbb{E}\left[ \hat{\theta}^T \Sigma W \cdot 1_{\mathcal{E}_p} \right]$$

$$\leq C_{\max} \mathbb{E}\left[ \|\Delta\|^2 \cdot 1_{\mathcal{E}_p} \right] + C_W \mathbb{E}\left[ \|\hat{\theta}\|^2 \cdot 1_{\mathcal{E}_p} \right].$$

We take the expectation with respect to $Z_\ast$ and $W_\ast$ to derive the second line. We proceed to upper bound the terms $\mathbb{E}[\|\Delta\|^2 1_{\mathcal{E}_p}]$ and $\mathbb{E}[\|\hat{\theta}\|^2 1_{\mathcal{E}_p}]$.

To study $\mathbb{E}[\|\hat{\theta}\|^2 1_{\mathcal{E}_p}]$, recalling that $\hat{M} = \hat{A}^T \hat{A}$, we have

$$\|\hat{\theta}\| = \|\hat{A}(\hat{A}^T X^T X A)^{-1} \hat{A}^T X^T Y\|$$

$$\leq \|\hat{A} \hat{M}^{-1}\|_{op} \|\hat{A}^T X^T X A \hat{M}^{-1}\|_{op} \|\hat{A}^T X^T Y\|$$

$$\leq \Lambda_{\min}^{-1/2} \|\hat{A}^T X^T X A \hat{M}^{-1}\|_{op}^{-1/2} \|\hat{A}^T X^T Y\| / \sqrt{n}. \quad (108)$$

The third line uses $\|(H^T H)^{1/2} H\|_{op}^2 = \|(H^T H)^{1/2}\|_{op}$ for any matrix $H$ and the last line holds on the event $\mathcal{E}_p$, by (i) of Lemma 28. We then upper bound the operator norm above. From Weyl’s inequality, (ii) of Lemma 27, and $\lambda_{\min}(M^{-1}A^T X^T X A M^{-1}) \gtrsim n C_{\min}$ on the event $\mathcal{E}_p$, we have

$$\frac{1}{\sqrt{n}} \sigma_{\min}(X \hat{A} M^{-1}) \geq \frac{1}{\sqrt{n}} \sigma_{\min}(X A M^{-1}) - \frac{1}{\sqrt{n}} \|X(\hat{A} M^{-1} - A M^{-1})\|_{op}$$

$$\gtrsim C_{\min}^{1/2} - C_{\max}^{1/2} R_M \gtrsim C_{\min}^{1/2} \quad (109)$$

where we use condition (106) to derive the last inequality. Since

$$\mathbb{E}[\|Y\|^2 / n] = \sigma^2 + \beta^T \Sigma \beta \leq \sigma^2 + \|\beta\|^2 C_{\max}, \quad (110)$$

we further have

$$C_W \mathbb{E}[\|\hat{\theta}\|^2 1_{\mathcal{E}_p}] \lesssim \frac{C_W}{\Lambda_{\min}} C_{\min}^{-1} (\sigma^2 + \|\beta\|^2 C_{\max}). \quad (111)$$
To upper bound $\mathbb{E}[\|\Delta\|^2 1_{E_p}]$, we use the result of Theorem 8. For any constant $a \in (0, 1)$, using $X = ZA^T + W$, the reverse triangle inequality, and the basic inequality $2xy \leq ax^2 + y^2/a$ yields

$$
\|Z\beta - X\hat{\theta}\|^2 \geq (\|Z\Delta\| - \|W\hat{\theta}\|)^2 \geq (1 - a)\|Z\Delta\|^2 + \left(1 - \frac{1}{a}\right)\|W\hat{\theta}\|^2.
$$

Thus, using that $\lambda_{\min}(Z^T Z) \gtrsim nC_{\min}$ on the event $E_p$,

$$
C_{\max}\mathbb{E}[\|\Delta\|^2 1_{E_p}] \lesssim \kappa(\Sigma^z) \frac{1}{n} \mathbb{E}\|Z\beta - X\hat{\theta}\|^2 + \kappa(\Sigma^z) \frac{1}{n} \mathbb{E}\|W\hat{\theta}\|^2 1_{E_p}.
$$

What remains is to upper bound the second term on the right hand side. On the event $E_p$, using (108) – (109), (iv) of Lemma 27 and (ii) of Lemma 28 gives

$$
\|W\hat{\theta}\| \leq \|W\hat{M}^{-1}\|_{op}(\hat{M}^{-1}A^T(X^T X/n)\hat{M}^{-1}A^T)^{1/2}\|Y\|/\sqrt{n}
\leq C_{\min}^{-1/2} (\|WAM^{-1}\|_{op} + \|W(\hat{M}^{-1}A - AM^{-1})\|_{op})\|Y\|/\sqrt{n}
\lesssim C_{\min}^{-1/2} \left(\Lambda_{\min}^{-1/2} C_W + R_M\right)\|Y\|.
$$

Therefore, by using (110) again, we obtain

$$
\frac{1}{n} \mathbb{E}\|W\hat{\theta}\|^2 1_{E_p} \lesssim C_{\min}^{-1}(\sigma^2 + \|\beta\|^2 C_{\max}) (\Lambda_{\min}^{-1} C_W + R_M^2).
$$

Finally, combining the result of Theorem 8 and (112) concludes

$$
C_{\max}\mathbb{E}[\|\Delta\|^2 1_{E_p}] \lesssim \kappa(\Sigma^z) \left[C_{\min}^{-1}(\sigma^2 + \|\beta\|^2 C_{\max}) \left(\frac{C_W}{\Lambda_{\min}} + R_M^2\right) + \frac{K}{n}\sigma^2\right]
$$

which together with (111) further yields

$$
\mathbb{E}\left[(Z^T\beta - \hat{\theta}^T X)\right]^2 1_{E_p}
\leq \kappa(\Sigma^z) \left[C_{\min}^{-1}(\sigma^2 + \|\beta\|^2 C_{\max}) \left(\frac{C_W}{\Lambda_{\min}} + R_M^2\right) + \frac{K}{n}\sigma^2\right]
$$

This completes the proof for $\mathbb{E}[(Z^T\beta - \hat{\theta})^2 1_{E_p}]$.

**B.12 Proof of lemmas 13, 14 and 15**

**B.12.1 Proof of Lemma 13**

Without loss of generality, we assume $\|v\| = 1$. Observe that

$$
\mathbb{E}\left[\exp(t\langle Z, v \rangle)\right] = \mathbb{E}\left[\exp\left(t\langle (\Sigma^z)^{-1/2} Z, (\Sigma^z)^{1/2} v \rangle\right)\right]
\leq \mathbb{E}\left[\exp\left(t^2\gamma^2 z v^T \Sigma^z v / 2\right)\right], \quad \forall t \in \mathbb{R}.
$$

This and the definition of the operator norm proves the statement of $\langle Z, v \rangle$. Choosing $v = e_k$ and using $\|\Sigma^z\|_{\infty} \leq B z$ conclude the proof of (1). Part (1) together with the independence between $Z$ and $W$ yields

$$
\mathbb{E}\left[\exp(tX)\right] = \mathbb{E}\left[\exp(t\langle A_j, Z \rangle)\right] \mathbb{E}\left[\exp(tW)\right]
\leq \exp\left(t^2\gamma^2 z A_j^T \Sigma^z A_j / 2\right) \exp\left(t^2\gamma^2 w / 2\right)
\leq \exp\left(t^2(\gamma^2 z \|\Sigma^z\|_{\infty} + \gamma^2 w / 2\right) \quad (\|A_j\|_1 \leq 1)
$$

for any $t \in \mathbb{R}$ and any $1 \leq j \leq p$. This completes the proof.
B.12.2 Proof of Lemma 14

By Assumption 2, $W_{tj}$ is $\gamma_w$ sub-Gaussian for all $t \in [n]$ and $j \in [p]$, so

$$E[\exp(tW_{ti})] \leq \exp(t^2 \gamma_w^2/2), \quad \forall t \in \mathbb{R}.$$ 

Again by Assumption 2, $W_{ti}$ is independent across index $i$, so using the previous display we find

$$E[\exp(t\eta_{tk})] = \prod_{i \in I} E[\exp(tW_{ti}/m_k)]$$

$$\leq \prod_{i \in I} \exp \left[ t^2 \gamma_w^2/(2m_k^2) \right]$$

$$\leq \exp \left[ t^2 \gamma_w^2/(2m) \right],$$

proving that $\eta_{tk}$ is $\gamma_w/\sqrt{m}$-subgaussian. Similarly, since $\eta_{tk}$ is independent across index $k$, we have $\sum_k v_k \eta_{tk}$ is $\gamma_w/\sqrt{m}$ by using $\|v\|^2 \leq 1$. This concludes the proof. 

B.12.3 Proof of Lemma 15

Let $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ denote, respectively, the sub-exponential norm and the sub-gaussian norm (Definitions 5.7 and 5.13 in Vershynin (2012)) as

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (E[|X|^p])^{1/p}, \quad \|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (E[|X|^p])^{1/p}$$

for any random variable $X$. Then, $\|X_t\|_{\psi_2} \leq c\gamma_x$ and $\|Y_t\|_{\psi_2} \leq c\gamma_y$ and an application of the Hölder’s inequality yields $\|X_tY_t\|_{\psi_1} \leq \|X_t\|_{\psi_2} \|Y_t\|_{\psi_2} \leq c\gamma_x\gamma_y$. The proof follows by Corollary 5.17 in Vershynin (2012). 

B.13 Proof of lemmas for proving Theorem 3

Before proving Lemma 11, we first give a useful fact.

**Fact 1.** Let $\Sigma = A\Sigma^z A^T + \Sigma^w$, $G = \Omega + A^T \Omega^w A$ with $\Omega = (\Sigma^z)^{-1}$ and $\Omega^w = (\Sigma^w)^{-1}$. 

$$\Sigma^{-1} A \Sigma^z = \Omega^w A G^{-1}, \quad \Sigma^z - \Sigma^z A^T \Sigma^{-1} A \Sigma^z = G^{-1}.$$ 

**Proof.** The Sherman-Morrison-Woodbury formula gives

$$\Sigma^{-1} A \Sigma^z = [\Omega^w - \Omega^w A(\Omega + A^T \Omega^w A)^{-1} A^T \Omega^w] A \Sigma^z$$

$$= \Omega^w A \Sigma^z - \Omega^w A(\Omega + A^T \Omega^w A)^{-1}(\Omega + A^T \Omega^w A - \Omega) \Sigma^z$$

$$= \Omega^w A(\Omega + A^T \Omega^w A)^{-1},$$

which concludes the proof of the first statement. The second part follows immediately by noting that 

$$\Sigma^z - \Sigma^z A^T \Sigma^{-1} A \Sigma^z = \Sigma^z - \Sigma^z A^T \Omega^w A(\Omega + A^T \Omega^w A)^{-1} = (\Omega + A^T \Omega^w A)^{-1}. $$

\[56\]
Proof of Lemma 11. By the additivity of the Kullback-Leibler divergence, it suffices to consider one data pair \((X_i, Y_i)\). We remove the subscript \(i\) for simplicity. Note that, for given \(\beta\) with \(j = 1, 2\), model (1) implies
\[
\begin{bmatrix}
Y \\
X
\end{bmatrix} \sim N_{p+1} \left( 0, \begin{bmatrix}
\beta^T \Sigma \beta_j + \sigma^2 & \beta_j^T \Sigma z A^T \\
A \Sigma z \beta_j & \Sigma
\end{bmatrix} \right)
\]
with \(\Sigma = A \Sigma z A^T + \Sigma^w\). This further yields
\[
Y|X \sim N(\beta^T \Sigma z A^T \Sigma^{-1} X, \sigma^2 + \beta_j^T (\Sigma z - \Sigma z A^T \Sigma^{-1} A \Sigma z) \beta_j)
\]
\[= N(\mu_j, \sigma^2_j). \tag{113} \]
Since the marginal distribution of \(X\) does not depend on \(\beta\), we observe that
\[
\frac{1}{n} \text{KL}(P_{\beta_1}, P_{\beta_2}) = E_{\beta_1} [\log f_{\beta_1}(Y|X)] - E_{\beta_1} [\log f_{\beta_2}(Y|X)]
\]
\[= -\frac{1}{2} \log \sigma_1^2 - \frac{1}{2\sigma_1^2} E_{\beta_1} [(Y - \mu_1)^2] + \frac{1}{2} \log \sigma_2^2 + \frac{1}{2\sigma_2^2} E_{\beta_1} [(Y - \mu_2)^2]
\]
\[= \frac{1}{2} (\log \sigma_2^2 - \log \sigma_1^2) + \frac{1}{2\sigma_2^2} E_{\beta_1} [(Y - \mu_2)^2 - (Y - \mu_1)^2] + \frac{\sigma_1^2 - \sigma_2^2}{2\sigma_2^2} \cdot E_{\beta_1} [(Y - \mu_1)^2]
\]
where the last inequality uses \(E_{\beta_1} [(Y - \mu_1)^2] = \sigma_1^2\) from (113). To calculate the expectation, it follows from \(E[X X^T] = \Sigma\) and (113) that
\[
E_{\beta_1} [(Y - \mu_2)^2 - (Y - \mu_1)^2] = E_{\beta_1} [\mu_2^2 - \mu_1^2 + 2Y(\mu_1 - \mu_2)]
\]
\[= \beta_2^T \Sigma z A^T \Sigma^{-1} A \Sigma z \beta_2 - \beta_1^T \Sigma z A^T \Sigma^{-1} A \Sigma z \beta_1 + 2E_{\beta_1} [Y(\beta_1 - \beta_2)^T \Sigma z A^T \Sigma^{-1} X]
\]
\[= \beta_2^T \Sigma z A^T \Sigma^{-1} A \Sigma z \beta_2 - \beta_1^T \Sigma z A^T \Sigma^{-1} A \Sigma z \beta_1 + 2E [(\beta_1 - \beta_2)^T \Sigma z A^T \Sigma^{-1} A Z T \beta_1]
\]
\[= (\beta_1 - \beta_2)^T \Sigma z A^T \Sigma^{-1} A \Sigma z (\beta_1 - \beta_2),
\]
where in the fourth line we use model (1). Plugging this into the KL-divergence yields
\[
\frac{1}{n} \text{KL}(P_{\beta_1}, P_{\beta_2}) = \frac{1}{2} (\log \sigma_2^2 - \log \sigma_1^2) + \frac{\sigma_1^2 - \sigma_2^2}{2\sigma_2^2}
\]
\[+ \frac{1}{2\sigma_2^2} (\beta_1 - \beta_2)^T \Sigma z A^T \Sigma^{-1} A \Sigma z (\beta_1 - \beta_2).
\]
Note that \(\Sigma z - \Sigma z A^T \Sigma^{-1} A \Sigma z = G^{-1}\) from Fact 1. Recalling that \(\sigma_j^2 = \sigma^2 + \beta_j^T G^{-1} \beta_j\) from (113) and Fact 1 and using the inequality
\[
|\log(\sigma^2 + t_2) - \log(\sigma^2 + t_1)| \leq \frac{|t_2 - t_1|}{\min(\sigma^2 + t_1, \sigma^2 + t_2)}
\]
57
for $t_1, t_2 > 0$ gives
\[
\frac{1}{n} \text{KL}(\mathbb{P}_{\beta_1}, \mathbb{P}_{\beta_2}) \leq \frac{|\beta_1^T G^{-1} \beta_1 - \beta_2^T G^{-1} \beta_2|}{\min(\sigma_1^2, \sigma_2^2)} + \frac{(\beta_1 - \beta_2)^T (\Sigma^z - G^{-1})(\beta_1 - \beta_2)}{2\sigma_2^2}.
\]
Using $\sigma_j^2 \geq \sigma^2 + \|\beta_j\|^2 \lambda_{\min}(G^{-1}) = \sigma^2 + \|\beta_j\|^2/\|G\|_op$ completes the proof.

\section*{B.14 Proof of lemmas for proving Theorem 4}

\subsection*{B.14.1 Proof of Lemma 17}

For any $a \in [K]$, by using $\hat{m}_a = m_a + |L_a|$, we have
\[
\hat{Z}_a = \frac{1}{m_a} \sum_{i \in \hat{I}_a} Z_{A_i} = \frac{m_a}{m_a} Z_a + \frac{1}{m_a} \sum_{i \in \hat{I}_a} Z_{A_i} = Z_a + \frac{1}{m_a} \sum_{i \in L_a} Z_{(A_i - e_a)}.
\]

The above display gives
\[
\frac{1}{n} u^T \bar{Z}^T \bar{W} v = \frac{1}{n} \sum_{a,b=1}^K u_a v_b \bar{Z}_a^T \bar{W}_b + \frac{1}{n} \sum_{a,b=1}^K u_a v_b \frac{1}{m_a \hat{m}_b} \sum_{i \in \hat{I}_a} (A_i - e_a)^T Z_i \sum_{j \in \hat{I}_b} W_j.
\]

We first bound $\Delta_2$ by
\[
|\Delta_2| \leq \|u\|_1 \max_{a \in [K]} \left| \frac{|L_a|}{m_a} \left( \frac{1}{n} (A_i - e_a)^T Z_i^T \sum_{b} \frac{v_b}{\hat{m}_b} \sum_{j \in \hat{I}_b} W_j \right) \right|
\leq \frac{8\delta}{\nu} \rho \|u\|_1 \|v\|_1 \left\{ \max_{a,b \in [K]} \left| \frac{1}{n} \sum_{t=1}^n Z_{ta} \left( \sum_{j \in L_b} W_{tj} + \sum_{j \in \hat{I}_b} W_{tj} \right) \right| \right\},
\]

where we use (68) and $|I_b| \leq \hat{m}_b$ in the last inequality. Recall that, for any $t \in [n]$ and $a \in [K]$, $Z_{ta}$ is $\gamma'_\nu$-subgaussian from Lemma 13. Since $\log(p \vee n) = o(n)$ guarantees that $\delta = o(1)$, applying Lemmas 14 - 15 and taking the union bounds over $a, b \in [K], j \in J_1^b$ (recall that $L_b \subseteq J_1^b$) give
\[
\mathbb{P} \left\{ |\Delta_2| \lesssim \gamma'_w \left( \frac{1}{\sqrt{m}} + \rho \|u\|_1 \|v\|_1 \sqrt{\frac{\log(p \vee n)}{n}} \right) \right\} \geq 1 - (p \vee n)^{-c}.
\]

58
Regarding $\Delta_1$, recall that

$$
\tilde{W}_b = \frac{1}{m_b} \left( \sum_{i \in I_b} W_{ti} + \sum_{i \in L_b} W_{ti} \right) = \frac{m_b}{m_b} \eta_b + \frac{1}{m_b} \sum_{i \in L_b} W_{ti}
$$

$$
= \eta_b + \frac{|L_b|}{m_b} \left( \frac{1}{|L_b|} \sum_{i \in L_b} W_{ti} - \eta_b \right)
$$

(118)

from $\tilde{I}_b = I_b \cup L_b$ and the definition of $\eta_{tk}$ in (68). This together with $\rho \geq \max_b |L_b|/\bar{m}_b$ gives

$$
|\Delta_1| \leq \frac{1}{n} \sum_{t=1}^{n} \sum_{b=1}^{K} v_b \eta_b \left| + \rho \|v\|_1 \max_{b \in [K]} \frac{1}{n} \sum_{t=1}^{n} \sum_{b=1}^{K} v_b \eta_b \right|
$$

$$
\leq \frac{1}{n} \sum_{t=1}^{n} \sum_{b=1}^{K} v_b \eta_b \left| + \rho \|u\|_1 \|v\|_1 \left\{ \max_{a,b \in [K]} \frac{1}{n} \sum_{t=1}^{n} Z_{ta} W_{ti} \right. + \max_{a,b \in [K]} \frac{1}{n} \sum_{t=1}^{n} Z_{ta} \eta_b \right\}.
$$

(119)

Take $u = e_k$ in (119) for some $k \in [K]$. Invoking Lemmas 13 – 15 and taking the union bounds yield

$$
P \left\{ \max_{k} \frac{1}{n} \sum_{b} v_b Z^T_k \tilde{W}_b \right\} \lesssim \gamma'_w \left( \frac{\|v\|}{\sqrt{m}} + \rho \|v\|_1 \right) \sqrt{\log(p \lor n)} \geq 1 - (p \lor n)^{-c}.
$$

(120)

Noting that $m \geq 2$ and $\rho \leq 1$, combining (117) with (120) concludes the proof of (95). Similarly, since $\langle z_t, u \rangle$ is $(\gamma_z \sqrt{u^T \Sigma z} u)$-subgaussian from Lemma 13, using Lemmas 13 – 15 again concludes

$$
\frac{1}{n} \sum_{a,b} u_a v_b Z^T_a \tilde{W}_b \lesssim \gamma_w \left( \gamma_z \frac{\|v\|}{\sqrt{m}} \sqrt{u^T \Sigma z} u \right) \sqrt{\log(p \lor n)}
$$

with probability greater than $1 - (p \lor n)^{-c}$. This and (117) concludes the proof of (96).

\[ \square \]

B.14.2 Proof of Lemma 18

Write $\tilde{W} := WA_I(A^T_J A_J)^{-1}$. For any $a \in [K]$, similar as (118), we write

$$
\tilde{W}_a = \frac{1}{m_a} \sum_{i \in I_a} W_i + \frac{1}{m_a} \sum_{i \in L_a} W_i = \tilde{W}_a + \frac{|L_a|}{m_a} \left( \frac{1}{|L_a|} \sum_{i \in L_a} W_i - \tilde{W}_a \right).
$$

(121)

59
Then we can expand the quadratic term as
\[
u^T \left( \frac{1}{n} \tilde{W}^T \tilde{W} - D \right) v = \sum_a u_a v_a \left( \frac{1}{n} \tilde{W}^T_a \tilde{W}_a - d_a \right) + \frac{1}{n} \sum_{a \neq b} u_a v_b (\tilde{W}_a^T \tilde{W}_b + \tilde{W}_a^T R_b + \tilde{W}_b^T R_a + R_a^T R_b) .
\]

By using \(X_i = ZA_i + W_i\) and the definition of \(d_a\) in (74), we observe that
\[
d_a = \frac{1}{m_a^2} \sum_{i \in \tilde{I}_a} \frac{1}{n} X_i^T X_i - \frac{1}{m_a^2 (m_a - 1)} \sum_{i \neq j \in \tilde{I}_a} \frac{1}{n} X_i^T X_j
\]
\[+ \frac{1}{m_a^2} \sum_{i \in \tilde{I}_a} \frac{1}{n} A_i^T Z^T Z A_i + \frac{1}{n} W_i^T W_i + \frac{2}{n} A_i^T Z^T W_i\]
\[+ \frac{1}{m_a^2 (m_a - 1)} \sum_{i \neq j \in \tilde{I}_a} \left[ \frac{1}{n} A_i^T Z^T Z A_j + \frac{1}{n} W_i^T W_j + \frac{2}{n} A_i^T Z^T W_j \right].
\]

Rearranging terms yields
\[
d_a = \frac{1}{m_a^2} \sum_{i,j \in \tilde{I}_a} \frac{1}{n} W_i^T W_j - \frac{1}{m_a (m_a - 1)} \sum_{i \neq j \in \tilde{I}_a} \frac{1}{n} W_i^T W_j
\]
\[+ \frac{1}{m_a^2} \sum_{i \in \tilde{I}_a} \frac{1}{n} A_i^T Z^T Z A_i - \frac{1}{m_a (m_a - 1)} \sum_{i \neq j \in \tilde{I}_a} \frac{1}{n} A_i^T Z^T Z A_j.
\]
\[+ \frac{2}{m_a^2} \sum_{i \in \tilde{I}_a} \frac{1}{n} A_i^T Z^T W_i - \frac{2}{m_a (m_a - 1)} \sum_{i \neq j \in \tilde{I}_a} \frac{1}{n} A_i^T Z^T W_j.
\]

Note the first term on the RHS is equal to \(n^{-1} \tilde{W}_a^T \tilde{W}_a\). We further obtain
\[
\left| u^T \left( n^{-1} \tilde{W}^T \tilde{W} - D \right) v \right| \leq \|u\| \|v\| \max_a (|T_1^a| + |T_2^a| + |T_3^a|)
\]
\[+ \frac{1}{n} \sum_{a \neq b} u_a v_b (\tilde{W}_a^T \tilde{W}_b + \tilde{W}_a^T R_b + \tilde{W}_b^T R_a + R_a^T R_b)
\]
\[:= \Delta_1 + \Delta_2
\]

**To bound \(\Delta_1\):** We first study \(T_1^a\), \(T_2^a\) and \(T_3^a\), separately. For \(T_1^a\), expanding \(\tilde{I}_a = I_a \cup L_a\) upper
bounds it by

\[
\frac{1}{\hat{m}_a(\hat{m}_a - 1)} \left\{ \left| \sum_{i \neq j \in I_a} \frac{1}{n} W_i^T W_j \right| + 2 \left| \sum_{i \in I_a, j \in L_a} \frac{1}{n} W_i^T W_j \right| + \left| \sum_{i \neq j \in L_a} \frac{1}{n} W_i^T W_j \right| \right\}
\leq \frac{1}{\hat{m}_a(\hat{m}_a - 1)} \left\{ \left| \sum_{i \neq j \in I_a} \frac{1}{n} W_i^T W_j \right| + \frac{2|I_a||L_a|}{\hat{m}_a(\hat{m}_a - 1)} \max_{j \in L_a} \left| \sum_{i \in I_a} \frac{1}{n} W_i^T W_j \right| \right.
\left. + \frac{|L_a||I_a|}{\hat{m}_a(\hat{m}_a - 1)} \max_{i \neq j \in I_a} \left| \frac{1}{n} W_i^T W_j \right| \right\}.
\]

For the first term, note that it is no greater than

\[
\frac{1}{n} \left| \sum_{i = 1}^n m_a \sum_{i \in I_a} W_{ti} \frac{1}{m_a - 1} \sum_{j \in I_a \setminus \{i\}} W_{tj} \right| = \frac{1}{n} \left| \sum_{i = 1}^n \eta_a \frac{1}{m_a - 1} \sum_{j \in I_a \setminus \{i\}} W_{tj} \right|.
\]

Since \((m_a - 1)^{-1} \sum_{j \in I_a \setminus \{i\}} W_{tj}\) is \((\gamma_w / \sqrt{m_a - 1})\)-subgaussian by the arguments of Lemma 14 and

\[
\mathbb{E} \left[ \sum_{i \in I_a} W_{ti} \sum_{j \neq i} W_{tj} \right] = 0,
\]

invoking Lemma 15 gives

\[
\mathbb{P} \left\{ \frac{1}{\hat{m}_a(\hat{m}_a - 1)} \left| \sum_{i \neq j \in I_a} \frac{1}{n} W_i^T W_j \right| \leq c_{\gamma_w}^2 \sqrt{\frac{\log(p \lor n)}{nm_a(m_a - 1)}} \right\} \geq 1 - (p \lor n)^{-c'}.
\]

Note that \(|L_a|/\hat{m}_a \leq \rho\) and \(\hat{m}_a - 1 \geq m_a\) when \(L_a \neq \emptyset\). We further obtain

\[
|T_{1a}^q| \leq \frac{1}{\hat{m}_a(\hat{m}_a - 1)} \left| \sum_{i \neq j \in I_a} \frac{1}{n} W_i^T W_j \right| + 2\rho \max_{j \in L_a} \frac{1}{n} \sum_{t = 1}^n W_{tj} \eta_a
\left. + \rho^2 \max_{i \neq j \in I_a} \left| \frac{1}{n} W_i^T W_j \right| \right\}.
\]

Invoking Lemma 15 and taking the union bound conclude

\[
\max_a |T_{1a}^q| \leq c_{\gamma_w}^2 \left( \frac{1}{\sqrt{m(m - 1)}} \lor \frac{\rho}{\sqrt{m}} \lor \rho^2 \right) \sqrt{\frac{\log(p \lor n)}{n}},
\] (123)
with probability greater than 1 \( - (p \lor n)^{-c} \). We bound \( T_2^a \) by writing

\[
T_2^a = \frac{m_a}{m_a^2} \frac{1}{n} Z_a^T Z_a + \frac{1}{m_a^2} \sum_{i \in L_a} \frac{1}{n} A_i^T Z_a^T Z A_i \]

\[
- \frac{1}{m_a^2 (m_a - 1)} \left\{ m_a (m_a - 1) \frac{1}{n} Z_a^T Z_a + 2 m_a \sum_{i \in L_a} \frac{1}{n} Z_a^T Z A_i \right\} \]

\[
+ \sum_{i \neq j \in L_a} \frac{1}{m_a^2 (m_a - 1)} \left\{ A_i^T Z_a^T Z A_j \right\} \]

\[
= \frac{1}{m_a^2} \sum_{i \in L_a} \left( \frac{1}{n} A_i^T Z_a^T Z A_i - \frac{1}{n} Z_a^T Z A_i \right) - \frac{2 m_a}{m_a^2 (m_a - 1)} \sum_{i \in L_a} \frac{1}{n} Z_a^T Z (A_i - e_a) \]

\[
- \frac{1}{m_a^2 (m_a - 1)} \sum_{i \neq j \in L_a} \left( A_i^T Z_a^T Z A_j - \frac{1}{n} Z_a^T Z_a \right). \]

Note that (69) and the event \( \mathcal{E} \) in (66) gives

\[
\left| \frac{1}{n} Z_a^T Z (A_i - e_a) \right| \leq \frac{8 \delta}{\nu} \max_{\ell} \left| \frac{1}{n} Z_a^T Z_{\ell} \right| \leq \frac{8 B_z}{\nu} \delta, \quad (124) \]

for any \( i \in L_b \) and \( a \in [K] \), which in conjunction with \( \|A_j\|_1 \leq 1 \) further gives

\[
\left| \frac{1}{n} A_i^T Z_a^T Z A_j - \frac{1}{n} Z_a^T Z_a \right| \leq \left| \frac{1}{n} (A_i - e_a)^T Z_a^T Z A_j \right| + \left| \frac{1}{n} Z_a^T Z (A_j^T - e_a) \right| \]

\[
\leq \frac{16 B_z}{\nu} \delta. \]

We thus obtain, on the event \( \mathcal{E} \),

\[
\max_a |T_2^a| \leq \max_a \left\{ \frac{|L_a|}{m_a^2} + \frac{m_a |L_a|}{m_a^2 (m_a - 1)} + \frac{|L_a| (|L_a| - 1)}{m_a^2 (m_a - 1)} \right\} \frac{16 B_z}{\nu} \delta \]

\[
\leq \frac{32 B_z \rho \delta}{\nu m} \leq \rho \sqrt{\frac{\log(p \lor n)}{nm^2}}. \quad (125) \]

Regarding \( T_3^a \), notice that

\[
|T_3^a| \leq \frac{2}{m_a^2} \sum_{i \in I_a} \frac{1}{n} |A_i^T Z_a^T W_i| + \frac{2}{m_a^2 (m_a - 1)} \sum_{i \neq j \in I_a} \frac{1}{n} |A_i^T Z_a^T W_j| \]

\[
\leq \frac{2}{m_a} \max_{i \in I_a} \frac{1}{n} \sum_{t=1}^{n} |\langle z_t, A_i \rangle W_{ti}| + \frac{2}{m_a} \max_{i \neq j \in I_a} \frac{1}{n} \sum_{t=1}^{n} |\langle z_t, A_i \rangle W_{tj}| \]

\[
\leq \frac{2}{m_a} \max_{a \in [K]} \frac{1}{n} \sum_{t=1}^{n} |Z_{tk} W_{ti}| + \frac{2}{m_a} \max_{a \in [K]} \frac{1}{n} \sum_{t=1}^{n} |Z_{tk} W_{tj}| \]

62
by using \( \|A_i\|_1 \leq 1 \) to derive the last inequality. Invoking Lemmas 14–15 and taking the union bounds over \( a \in [K], i, j \in I_k \cup J_1^k \) yield

\[
P \left\{ \max_a |T_3^a| \lesssim \gamma' w \sqrt{\frac{\log(p \wedge n)}{nm^2}} \right\} \geq 1 - (p \wedge n)^{-c}. \tag{126}
\]

Collecting displays (123) - (126) concludes

\[
\Delta_1 \lesssim \|u\|_1 \|v\|_w \gamma' z \left( \frac{1}{\sqrt{m(m-1)}} \vee \frac{\rho}{\sqrt{m}} \vee \rho^2 \right) \sqrt{\frac{\log(p \wedge n)}{n}}, \tag{127}
\]

with probability greater than \( 1 - (p \vee n)^{-c} \).

**To bound \( \Delta_2 \):** We study the first term

\[
\frac{1}{n} \sum_{a \neq b} u_a v_b W_a^T W_b = \frac{1}{n} \sum_{a=1}^n \sum_{b \neq a} u_a \eta_a \sum_{b \neq a} v_b \eta_b.
\]

Since Lemma 14 guarantees \( \sum_{a} u_a \eta_a \) is \( (\|u\|_w / \sqrt{m}) \)-subgaussian and, similarly, \( \sum_{b \neq a} v_b \eta_b \) is \( (\|v\|_w / \sqrt{m}) \)-subgaussian, invoking Lemma 15 and noting that

\[
E \left[ \sum_a u_a \eta_a \sum_{b \neq a} v_b \eta_b \right] = \sum_{a \neq b} u_a v_b E[\eta_a \eta_b] = 0
\]

gives

\[
P \left\{ \frac{1}{n} \sum_{a \neq b} u_a v_b W_a^T W_b \leq \|u\|_1 \|v\|_w \gamma^2 \sqrt{\frac{\log(p \wedge n)}{nm^2}} \right\} \geq 1 - (p \wedge n)^{-c}. \tag{128}
\]

We then bound the rest term by term. By recalling that (121), we obtain

\[
\frac{1}{n} \left| \sum_{a \neq b} u_a v_b W_a^T R_b \right| \\
= \frac{2}{n} \sum_{b=1}^K \sum_{t=1}^n \left| \sum_{b \neq a} v_b \sum_{\ell=1}^t \frac{L_b \eta_{b\ell} - \sum_{i \in L_b} W_{ti}}{|L_b|} \sum_{a \neq b} u_a \eta_a \right| \\
\leq 2 \|v\|_1 \max_{b \in [K]} \left| \frac{L_b}{m_b} \right| \frac{1}{n} \sum_{t=1}^n \left| \sum_{a \neq b} u_a \eta_a \right| + 2 \|v\|_1 \max_{b \in [K]} \left| \frac{1}{n} \sum_{t=1}^n W_{ti} \right| \frac{1}{n} \sum_{t=1}^n \left| \sum_{a \neq b} u_a \eta_a \right|.
\]

Invoking Lemmas 14 and 15 and taking union bounds conclude

\[
P \left\{ \frac{1}{n} \left| \sum_{a \neq b} u_a v_b W_a^T R_b \right| \leq C p \|u\|_1 \|v\|_w^2 \sqrt{\frac{\log(p \wedge n)}{nm}} \right\} \geq 1 - (p \vee n)^{-c}. \tag{129}
\]
A similar bound can be obtained for $|\sum_{a \neq b} u_a v_b \tilde{W}_b^T R_a|$. Finally, to bound the last term, we have

$$
\frac{1}{n} \left| \sum_{a \neq b} u_a v_b \tilde{R}_b^T R_a \right|
= \frac{1}{n} \left| \sum_{a \neq b} \frac{u_a v_b}{m_b \eta_{b_a}} \sum_{t=1}^{n} \left( |L_b| \eta_{b_t} - \sum_{i \in L_b} W_{ti} \right) \left( |L_a| \eta_{a_t} - \sum_{i \in L_a} W_{ti} \right) \right|.
$$

Thus, it can be upper bounded by

$$
\rho^2 \|u\|_1 \|v\|_1 \max_{a \neq b} \left| \frac{1}{n} \sum_{t=1}^{n} \left( \eta_{b_t} - \frac{1}{|L_b|} \sum_{i \in L_b} W_{ti} \right) \left( \eta_{a_t} - \frac{1}{|L_a|} \sum_{i \in L_a} W_{ti} \right) \right| 
\leq \rho^2 \|u\|_1 \|v\|_1 \max_{a \neq b} \left| \frac{1}{n} \sum_{t=1}^{n} \eta_{b_t} \eta_{a_t} \right| + 2 \rho^2 \|u\|_1 \|v\|_1 \max_{a \neq b \in [K]} \left| \frac{1}{n} \sum_{t=1}^{n} W_{ti} \eta_{a_t} \right| 
+ \rho^2 \|u\|_1 \|v\|_1 \max_{a \neq b \in [K]} \left| \frac{1}{n} \sum_{t=1}^{n} W_{ti} W_{tj} \right|.
$$

Invoking Lemmas 14 and 15 and taking the union bounds yield

$$
P \left\{ \frac{1}{n} \left| \sum_{a \neq b} u_a v_b \tilde{R}_b^T R_a \right| \leq \gamma_w^2 \rho^2 \|u\|_1 \|v\|_1 \sqrt{\log(p \vee n)} \right\} \geq 1 - (p \vee n)^{-c}.
$$

Combining (128), (129) and (131) concludes

$$
P \left\{ \Delta_2 \leq \gamma_w^2 \left( \frac{\|u\|}{\sqrt{m}} + \rho \|u\|_1 \right) \left( \frac{\|v\|}{\sqrt{m}} + \rho \|v\|_1 \right) \sqrt{\log(p \vee n)} \right\} \geq 1 - (p \vee n)^{-c}.
$$

Displays (127) and (132) complete the proof.

**B.14.3 Proof of Lemma 19**

From (67), by noting that

$$Z - \tilde{Z} = Z - Z A_f^T \hat{A}_f (\hat{A}_f^T \hat{A}_f)^{-1} = Z \left[ I - A_f^T \hat{A}_f (\hat{A}_f^T \hat{A}_f)^{-1} \right] := Z \Delta_{\hat{f}}
$$

we obtain

$$
\frac{1}{n} \|X^T (Z - \tilde{Z}) \beta\|_{\infty} \leq \frac{1}{n} \|X^T Z\|_{\infty} \|\Delta_{\hat{f}} \beta\|_1 \leq \frac{1}{n} \|X^T Z\|_{\infty} \|\Delta_{\hat{f}}\|_{1,\infty} \|\beta\|_1.
$$
For any $k, \ell \in [K]$, we have
\[
\frac{1}{n} \hat{X}_{\ell}^T Z_{\ell} = \frac{1}{m_k} \sum_{i \in \hat{I}_k} \frac{1}{n} X_i^T Z_{\ell} \leq \max_{i \in \hat{I}_k} \frac{1}{n} |X_i^T Z_{\ell}|
\]
\[
\leq \max_{i \in \hat{I}_k} \frac{1}{n} \sum_{t=1}^{n} (A_i^T z_t) Z_{\ell t} + \max_{i \in \hat{I}_k} \frac{1}{n} |W_i^T Z_{\ell}| \quad \text{(since $X_i = ZA_i + W_i$)}
\]
\[
\leq \max_{k \in [K]} \frac{1}{n} \sum_{t=1}^{n} Z_{ik} Z_{\ell t} + \max_{i \in \hat{I}_k} \frac{1}{n} |W_i^T Z_{\ell}|. \quad \text{(since $\|A_i\| \leq 1$)}
\]

Since $W_i$ is $\gamma_w$-subgaussian and $Z_{\ell t}$ is $\gamma'_w$-subgaussian from Lemma 13, invoking the event $E$ in (66), Lemma 15 and taking the union bounds yield
\[
\frac{1}{n} \|\hat{X}^T Z\|_\infty \lesssim B_z + \gamma'_w \sqrt{\frac{\log(p \vee n)}{n}} = O(1) \tag{134}
\]
with probability greater than $1 - (p \vee n)^{-c}$, where we use $\log p = o(n)$ and $B_z = O(1)$ in the last inequality. On the other hand, observing that
\[
\Delta_{\hat{f}} = (\hat{A}_{\hat{f}} - A_{\hat{f}})^T \hat{A}_{\hat{f}} \cdot \text{diag} (\hat{\gamma}_1^{-1}, \ldots, \hat{\gamma}_K^{-1})
\]
and $\hat{A}_{ik} = 1$ for any $i \in \hat{I}_k$ and $k \in [K]$, we have
\[
\|\Delta_{\hat{f}}\|_{1,\infty} = \max_{k \in [K]} \frac{1}{m_k} \sum_{\ell=1}^{K} \sum_{i \in \hat{I}} (\hat{A}_{i\ell} - A_{i\ell}) \hat{A}_{ik}
\]
\[
= \max_{k \in [K]} \frac{1}{m_k} \sum_{\ell=1}^{K} \sum_{i \in \hat{I}_k} (\hat{A}_{i\ell} - A_{i\ell}) \hat{A}_{ik} \quad \text{(since $\hat{A}_{ik} = 0$, $\forall i \in \hat{I} \setminus \hat{I}_k$)}
\]
\[
= \max_{k \in [K]} \frac{1}{m_k} \sum_{\ell=1}^{K} \sum_{i \in \hat{I}_k} (\hat{A}_{i\ell} - A_{i\ell}) \hat{A}_{ik} \quad \text{(since $\hat{A}_{i\ell} = A_{i\ell}$, $\forall i \in I_k$)}
\]
\[
\leq \max_{k \in [K]} \frac{|L_k|}{m_k} \max_{i \in L_k} \|\hat{A}_{i\ell} - A_{i\ell}\|_1.
\]
Using the definition of $\rho$ and (69) concludes, on the event $\mathcal{E}$,
\[
\|\Delta_{\hat{f}}\|_{1,\infty} \lesssim \rho \sqrt{\log(p \vee n)/n}. \tag{135}
\]
Combining this with (134) completes the proof. \hfill \qed

B.14.4 Proof of Lemma 20

We work on the event $\mathcal{E}$ defined in (66) which implies $\hat{K} = K$. Let $\hat{\mathcal{I}} = \{\hat{I}_1, \ldots, \hat{I}_K\}$ be the estimated partition of pure variables. Recall that, from the proof of Theorem 4, $\hat{\Sigma} = n^{-1} \hat{X}^T \hat{X} - D$ with
\[ D = \text{diag}(d_1, \ldots, d_K) \text{ and } d_k \text{ defined in (74). By } \tilde{X} = \tilde{Z} + \tilde{W}, \text{ we obtain} \]

\[
u^T(\tilde{\Sigma}^z - \Sigma^z)v = \nu^T \left( \frac{1}{n} \tilde{X}^T \tilde{X} - \Sigma^z \right)v - \nu^T Dv = \nu^T \left( \frac{1}{n} \tilde{Z}^T \tilde{Z} - \Sigma^z \right)v + \nu^T \left( \frac{1}{n} \tilde{W}^T \tilde{W} - D \right)v + \frac{1}{n} \nu^T \tilde{W}v + \frac{1}{n} \nu^T \tilde{W}u \]

\[ := T_1 + T_2 + T_3 + T_4. \]

Plugging (114) into \( T_1 \) yields

\[
T_1 = \nu^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right)v + \sum_{a,b} u_a v_b \left[ \frac{1}{\tilde{m}_b} \sum_{i \in \mathcal{L}_b} \frac{1}{n} Z_a^T Z(A_i - e_b) \right] + \frac{1}{\tilde{m}_a} \sum_{i \in \mathcal{L}_a} \frac{1}{n} Z_b^T Z(A_i - e_a) + \frac{1}{\tilde{m}_a \tilde{m}_b} \sum_{i \in \mathcal{L}_a, j \in \mathcal{L}_b} \frac{1}{n} (A_i - e_a)^T Z^T Z(A_j - e_b) \]

\[ := \nu^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right)v + T_{11}. \quad (136) \]

Since Lemma 13 guarantees that \( \langle z_t, v \rangle \) and \( \langle z_t, u \rangle \) are \( (\gamma_z \sqrt{u^T \Sigma^z v}) \)-subgaussian and \( (\gamma_z \sqrt{u^T \Sigma^z u}) \)-subgaussian, respectively, by recalling that \( n^{-1} \sum[Z^T Z] = \Sigma^z \) and invoking Lemma 15 give

\[
\mathbb{P} \left\{ \left| \nu^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right)v \right| \leq \sqrt{u^T \Sigma^z u \sqrt{v^T \Sigma^z v} \sqrt{\log(p \land n)}} \right\} \geq 1 - (p \land n)^{-c}. \quad (137) \]

To bound \( T_{11} \), for any \( i \in \mathcal{L}_a, j \in \mathcal{L}_b \) and \( a, b \in [K] \), using (124) and \( \|A_i\|_1 \leq 1 \) yields

\[
\left| \frac{1}{n} (A_i - e_a)^T Z^T Z(A_j - e_b) \right| \leq \frac{1}{n} A_i^T Z^T Z(A_j - e_b) + \frac{1}{n} Z_a^T Z(A_j - e_b) \leq \frac{16 B}{\nu} \delta. \quad (138) \]

Recall that \( \max_a |\mathcal{L}_a| / \tilde{m}_a \leq \rho \). Displays (124) and (138) imply

\[
|T_{11}| \leq \|u\|_1 \|v\|_1 \max_{a,b} \left( \frac{|L_b|}{\tilde{m}_b} + \frac{|L_a|}{\tilde{m}_a} + \frac{2|L_a||L_b|}{\tilde{m}_a \tilde{m}_b} \right) \frac{8B}{\nu} \delta \]

\[ \leq \rho \|u\|_1 \|v\|_1 \sqrt{\log(p \land n)} \frac{1}{n}. \quad (139) \]

Therefore, together with the result in (137), we conclude

\[
\mathbb{P} \left\{ |T_1| \leq c \left( \gamma_z^2 \sqrt{u^T \Sigma^z u \sqrt{v^T \Sigma^z v}} \lor \rho \|u\|_1 \|v\|_1 \right) \sqrt{\log(p \land n)} \frac{1}{n} \right\} \geq 1 - (p \land n)^{-c}. \quad (140) \]

The proof is completed by invoking Lemmas 17 – 18 for \( T_2 - T_4 \) and using \( u^T \Sigma^z u \geq C_{\min} \|u\|^2 \) to simplify expressions.
B.14.5 Proof of Lemma 21

To show the upper bound of \(\|n^{-1} Z^T Z - \Sigma^z\|_{op}\), in light of (137), we use the classical discretization method to prove the uniformity over \(v \in S^{K-1}\). Let \(\mathcal{N}_\varepsilon \subset S^{K-1}\) be a minimal \(\varepsilon\)-net of \(S^{K-1}\), i.e. a set with minimum cardinality such that the collection of \(\varepsilon\)-balls centered at points in \(\mathcal{N}_\varepsilon\) covers \(S^{K-1}\). From Lemma 5.4 in Vershynin (2012), for any \(\varepsilon \in (0, 1/2)\), we know

\[
\sup_{v \in S^{K-1}} \left| v^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right) v \right| \leq (1 - 2\varepsilon)^{-1} \max_{v \in \mathcal{N}_\varepsilon} \left| v^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right) v \right|.
\]

Taking \(\varepsilon = 1/3\) yields

\[
\mathbb{P} \left\{ \sup_{v \in S^{K-1}} \left| v^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right) v \right| \geq t \right\} \leq |\mathcal{N}_{1/3}| \cdot \mathbb{P} \left\{ \left| v^* T \left( \frac{1}{n} Z^T Z - \Sigma^z \right) v^* \right| \geq \frac{t}{3} \right\},
\]

for some \(v^* \in \mathcal{N}_{1/3}\). By using \(|\mathcal{N}_{1/3}| \leq 7^K\) from Lemma 5.2 in Vershynin (2012), choosing \(u = v = v^*\) in (137) together with the definition of \(\lambda_1\), we obtain

\[
|\mathcal{N}_{1/3}| \cdot \mathbb{P} \left\{ \left| v^* T \left( \frac{1}{n} Z^T Z - \Sigma^z \right) v^* \right| \geq c C_{\text{max}} \sqrt{\frac{K \log(p \vee n)}{n}} \right\} \leq (p \vee n)^{-c'} K.
\]

This completes the proof of the upper bound on \(\|n^{-1} Z^T Z - \Sigma^z\|_{op}\).

To prove the upper bound on \(\|\hat{\Sigma}^z - \Sigma^z\|_{op}\), as in the proof of Lemma 20, we consider the terms \(T_1 - T_4\) separately (except here \(T_3 = T_4\) since \(u = v\)). Specifically, we will upper bound

\[
T'_1 + T'_2 + T'_3 := \sup_{v \in S^{K-1}} \left| v^T \left( \frac{1}{n} Z^T \hat{Z} - \Sigma^z \right) v \right| + \sup_{v \in S^{K-1}} \left| v^T \left( \frac{1}{n} \tilde{W}^T \tilde{W} - D \right) v \right| + \sup_{v \in S^{K-1}} \left| \frac{2}{n} v^T \tilde{Z}^T \tilde{W} v \right|.
\]

For \(T'_1\), (136) implies

\[
T'_1 \leq \sup_{v \in S^{K-1}} \left| v^T \left( \frac{1}{n} Z^T Z - \Sigma^z \right) v \right| + \sup_{v \in S^{K-1}} |T_{11}|
\]

\[
\overset{(139)}{\leq} \left\| \frac{1}{n} Z^T Z - \Sigma^z \right\|_{op} + c \rho K \sqrt{\frac{\log(p \vee n)}{n}},
\]

by using \(\|v\|_2^2 \leq K\), where the last inequality holds on the event \(\mathcal{E}\) defined in (66). Invoking the above result on \(\|n^{-1} Z^T Z - \Sigma^z\|_{op}\) and the fact that \(\mathbb{P}(\mathcal{E}) \geq 1 - (p \vee n)^{-c'}\), we find that

\[
\mathbb{P} \left\{ T'_1 \leq c \left( C_{\text{max}} \vee \rho \sqrt{K} \right) \sqrt{\frac{K \log(p \vee n)}{n}} \right\} \geq 1 - (p \vee n)^{-c'}.
\]
To bound $T'_2$, from (122), it suffices to bound $\sup_{v \in \mathcal{S}^{K-1}} \Delta_2$ since the bound of $\Delta_1$ is uniformly in $v$. Display (122) gives

$$
\sup_{v \in \mathcal{S}^{K-1}} \Delta_2 = \sup_{v \in \mathcal{S}^{K-1}} \frac{1}{n} \left| \sum_{a \neq b} v_a v_b W_a^T \tilde{W}_b \right| + \sup_{v \in \mathcal{S}^{K-1}} \frac{1}{n} \left| \sum_{a \neq b} v_a v_b \left( 2 W_a^T R_b + R_a^T R_b \right) \right|.
$$

By repeating a discretization argument similar to the above, one can show from (128) that

$$
\mathbb{P} \left\{ \sup_{v \in \mathcal{S}^{K-1}} \frac{1}{n} \left| \sum_{a \neq b} v_a v_b W_a^T \tilde{W}_b \right| \leq c \sqrt{\frac{K \log(p \vee n)}{nm^2}} \right\} \geq 1 - (p \vee n)^{-c' K} \tag{142}
$$

and that, from (129) with $\|v\|_1 \leq \sqrt{K}$ uniformly,

$$
\mathbb{P} \left\{ \sup_{v \in \mathcal{S}^{K-1}} \frac{2}{n} \left| \sum_{a \neq b} v_a v_b W_a^T R_b \right| \leq c p K \sqrt{\frac{\log(p \vee n)}{nm}} \right\} \geq 1 - (p \vee n)^{-c' K}. \tag{143}
$$

From (130) by using $\|v\|_1^2 \leq K$, the last term $\frac{1}{n} \sum v_a v_b R_a^T R_b$ can be upper bounded by

$$
\sup_{v \in \mathcal{S}^{K-1}} \frac{1}{n} \left| \sum_{a \neq b} v_a v_b R_a^T R_b \right| \leq \rho^2 K \max_{a \neq b} \left\{ \frac{1}{n} \sum_{t=1}^n \eta_{tb} \eta_{ta} + \rho^2 K \max_{i \in L_a} \frac{1}{n} \sum_{t=1}^n W_{ti} \eta_{ta} \right\}
$$

$$
+ \rho^2 K \max_{i \in L_a} \left\{ \frac{1}{n} \sum_{t=1}^n W_{ti} \eta_{tb} + \rho^2 K \max_{a \neq b \in [K]} \frac{1}{n} \sum_{t=1}^n W_{tb} W_{tj} \right\}.
$$

Invoking Lemmas 14 - 15 and taking the union bound conclude that

$$
\mathbb{P} \left\{ \sup_{v \in \mathcal{S}^{K-1}} \frac{1}{n} \left| \sum_{a \neq b} v_a v_b R_a^T R_b \right| \leq c \rho^2 K \sqrt{\frac{\log(p \vee n)}{n}} \right\} \geq 1 - (p \vee n)^{-c'}. \tag{144}
$$

Finally, collecting (142) - (144) and invoking the bound of $\Delta_1$ via (123), (125) and (126) yield

$$
\mathbb{P} \left\{ T'_2 \leq c \left( \frac{1}{m} \vee \rho^2 \sqrt{K} \right) \sqrt{\frac{K \log(p \vee n)}{n}} \right\} \geq 1 - (p \vee n)^{-c'}. \tag{145}
$$

We then proceed to bound $T'_3$. From (115), (116) and (119), by using $\|v\|_1 \leq \sqrt{K}$, $\rho \leq 1$ and $\delta = o(1)$, we have

$$
T'_3 \leq \sup_{v \in \mathcal{S}^{K-1}} \frac{2}{n} \left| \sum_{t=1}^n \langle z_t, v \rangle \sum_{b} v_b \eta_{tb} \right| + \rho \sqrt{K} \max_{a,b \in [K]} \frac{2}{n} \sum_{t=1}^n Z_{ta} W_{ti}
$$

$$
+ \rho K \max_{a,b \in [K]} \frac{1}{n} \left| \sum_{t=1}^n Z_{ta} \eta_{tb} \right|.
$$
The discretization arguments together with Lemmas 13 - 15 give
\[
P \left\{ T'_3 \leq c \left( \sqrt{\frac{C_{\text{max}}}{m}} \lor \rho \lor \frac{K}{m} \right) \sqrt{\frac{K \log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-\epsilon}. \tag{146}\]
Collecting the bounds for \( T'_1, T'_2 \) and \( T'_3 \) completes the proof. \hfill \box

**B.14.6 Proof of Lemma 22**

We work on the event \( E \cap E_1 \) such that \( \hat{\Omega} = (\hat{\Sigma}^2)^{-1} \) exists and \( \|\hat{\Omega}\|_{op} = \lambda_{\text{min}}^{-1}(\hat{\Sigma}^2) = O(C_{\text{min}}^{-1}) \) (see the proof of Theorem 4). Note that it suffices to upper bound
\[
\|\hat{\Omega}\|_{\infty,1} \cdot \frac{1}{n} \left\| \frac{1}{n}(\tilde{X} - \hat{X})^T Y \right\|_{\infty} + \|\hat{\Omega}_k\|_{\infty,1} \left\| (\hat{\Sigma}^2 - \hat{\Sigma}^2) \beta \right\|_{\infty}.
\]
First, we have \( \|\hat{\Omega}\|_{\infty,1} \leq \sqrt{K} \|\hat{\Omega}\|_{op} = O(\sqrt{K}C_{\text{min}}^{-1}). \) Since \( \tilde{X} = \hat{Z} + \hat{W} \) and \( \hat{X} = Z + \hat{W} \), we have
\[
\frac{1}{n} \left\| \frac{1}{n}(\tilde{X} - \hat{X})^T Y \right\|_{\infty} \leq \frac{1}{n} \left\| (\hat{Z} - Z)^T Y \right\|_{\infty} + \frac{1}{n} \left\| (\hat{W} - W)^T Y \right\|_{\infty}.
\]
Fix any \( k \in [K] \). Display (114) yields
\[
\frac{1}{n} \left\| (\hat{Z}_k - Z_k)^T Y \right\| = \frac{1}{m_k} \left| \sum_{i \in L_k} (A_k - e_k)^T \frac{1}{n} Z^T Y \right| \leq \left\| \hat{L}_k \right\|_{\max} \left\| A_i - e_k \right\|_{\max} \frac{1}{n} \left\| Z^T Y \right\|.
\]
Since Lemma 15 together with \( \|Y_i\|_{\psi_2} = O(1 \lor \|\beta\|) \) for \( 1 \leq i \leq n \) and \( \log p = o(n) \) gives, with probability greater than \( 1 - (p \lor n)^{-\epsilon} \),
\[
\frac{1}{n} \max_{a \in [K]} \left| Z^T Y \right| \lesssim \|\Sigma^2 \beta\|_{\infty} + (1 \lor \|\beta\|) \sqrt{\frac{\log(p \lor n)}{n}} \lesssim C_{\text{max}}(1 \lor \|\beta\|),
\]
using \( \rho \geq |L_k|/\hat{m}_k \) and (69) concludes
\[
P \left\{ \frac{1}{n} \left| (\hat{Z}_k - Z_k)^T Y \right| \lesssim C_{\text{max}}(1 \lor \|\beta\|) \rho \sqrt{\frac{\log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-\epsilon}. \tag{147}\]
Similarly, using (121) gives
\[
\frac{1}{n} \left| (\hat{W} - W_k)^T Y \right| \leq \rho \max_{i \in J_k} \frac{1}{n} |W_i^T Y| + \rho \frac{1}{n} |\hat{W}_k^T Y|.
\]
Lemma 15 thus yields
\[
P \left\{ \frac{1}{n} \left| (\hat{W}_k - \hat{W}_k)^T Y \right| \lesssim C_{\text{max}}(1 \lor \|\beta\|) \rho \sqrt{\frac{\log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-\epsilon}. \tag{148}\]
Combining (147) and (148) concludes the rate of \( n^{-1}\| (\bar{X} - X)^T Y \|_\infty \).

It remains to upper bound \( \| (\hat{\Sigma} - \Sigma)^\beta \|_\infty \). By taking \( u = e_k, v = \beta \) in Lemmas 20, 17, 18 and inspecting their proofs, one can derive

\[
P \left\{ \| (\hat{\Sigma} - \Sigma)^\beta \|_\infty \lesssim \rho \| \beta \|_1 \sqrt{\frac{\log(p \lor n)}{n}} \right\}.
\]

In conjunction with the previous displays and \( \| \beta \|_1 \leq \sqrt{K} \| \beta \|_2 \), we conclude the proof. \( \square \)

B.15 Proof of lemmas for proving Proposition 7

B.15.1 Proof of Lemma 23

We work on the event \( E \cap E_1 \) defined in (66) and (70). To show the rate of \( \tau_i^2 - \tau_i^2 \), fix any \( i \in [p] \). Using \( X_i = ZA_i + W_i \) gives

\[
|\tau_i^2 - \tau_i^2| \leq \left| \frac{1}{n} A_i^T Z^T Z A_i - \hat{A}_i^T \hat{\Sigma} \hat{A}_i \right| + \frac{1}{n} W_i^T W_i - \tau_i^2 + \frac{2}{n} A_i^T Z^T W_i
\]

Since \( E[n^{-1} W_i^T W_i] = \tau_i^2 \) and \( |A_i^T Z^T W_i| \leq \| Z^T W_i \|_\infty \) by \( \| A_i \|_1 \leq 1 \), invoking Lemma 15 and taking the union bound yield

\[
P \left\{ \left| \frac{1}{n} W_i^T W_i - \tau_i^2 \right| + \frac{2}{n} A_i^T Z^T W_i \lesssim \sqrt{\frac{\log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c}.
\]

(149)

It remains to bound the first term. Note that, for any \( i \in I_k \) and \( k \in [K] \), \( \hat{A}_i = A_i \). Hence

\[
\left| \frac{1}{n} A_i^T Z^T Z A_i - \hat{A}_i^T \hat{\Sigma} \hat{A}_i \right| = \left| \frac{1}{n} \hat{Z}_k Z_k - \Sigma_{kk} \right|.
\]

Since \( Z_{kk} \) is \( \gamma'_z \)-subgaussian from Lemma 13, invoking Lemma 15 yields

\[
P \left\{ \left| \frac{1}{n} \hat{Z}_k Z_k - \Sigma_{kk} \right| \lesssim \gamma'_z \sqrt{\frac{\log(p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c}.
\]

This proves the rate of \( \tau_i^2 - \tau_i^2 \) for \( i \in I \). For any \( i \in J \),

\[
\left| \frac{1}{n} A_i^T Z^T Z A_i - \hat{A}_i^T \hat{\Sigma} \hat{A}_i \right| \leq \left| (\hat{A}_i - A_i)^T \hat{\Sigma} \hat{A}_i \right| + \left| (\hat{A}_i - A_i)^T \hat{\Sigma} A_i \right| + \left| A_i^T \left( \frac{1}{n} Z^T Z - \hat{\Sigma} \right) A_i \right|.
\]

Using Cauchy Schwarz inequality, \( \| \hat{A}_i \|_1 \leq 1 \) and Lemma 25 gives

\[
\left| (\hat{A}_i - A_i)^T \hat{\Sigma} \hat{A}_i \right| \leq \| \hat{\Sigma}(\hat{A}_i - A_i) \|_\infty = O_p(\sqrt{\log(p \lor n)/n}).
\]

The same bound can be obtained for the second term by noting \( \| A_i \|_1 \leq 1 \). Regarding the third term, using \( \| \hat{A}_i \|_1 \leq 1 \) again with (137) and Lemma 20 yields

\[
\| n^{-1} Z^T Z - \hat{\Sigma} \|_\infty \leq \| n^{-1} Z^T Z - \Sigma \|_\infty + \| \hat{\Sigma} - \Sigma \|_\infty = O_p(\sqrt{\log(p \lor n)/n}).
\]

70
Collecting (149) and the bounds of $R_1 - R_3$ concludes the proof of $|\bar{\tau}_i^2 - \tau_i^2|$ for any $i \in J$. This bound and $\log p = o(n)$ imply

$$\mathbb{P}\{\bar{\tau}_i^2 \geq \tau_i^2 - o(1)\} \geq 1 - (p \lor n)^{-c}.$$  

The proof of $|\bar{\tau}_i^{-2} - \tau_i^{-2}|$ follows by using the bound of $|\bar{\tau}_i^2 - \tau_i^2|$, $\tau_i^2 \geq C_W$ and noting that

$$|\bar{\tau}_i^{-2} - \tau_i^{-2}| = \frac{|\bar{\tau}_i^2 - \tau_i^2|}{\tau_i^2 \tau_i^2}.$$  

Taking the union bounds over $1 \leq i \leq p$ completes the proof.  

\[\Box\]

**B.15.2 Proof of Lemma 24**

We work on the event that the results of Theorems 2 and 4 hold intersecting with $\mathcal{E} \cap \mathcal{E}_1$ defined in (66) and (70). Thus, $\hat{\Sigma}$ is invertible. Recall that, $\hat{h}$ defined in (13) is equivalent to

$$\hat{h} = \frac{1}{n}(\hat{A}_i^T \hat{A}_i)^{-1} \hat{A}_i^T X_i^T Y \overset{(67)}{=} \frac{1}{n} \tilde{X}^T Y.$$  

Further, recall that $\tilde{\Sigma} = n^{-1} \tilde{X}^T \tilde{X} - D$ with $D = \text{diag}(d_1, \ldots, d_K)$ defined in (74). Using $Y = Z \beta + \varepsilon$ and $\tilde{X} = \tilde{Z} + \tilde{W}$ gives

$$\hat{\sigma}^2 = \frac{1}{n} Y^T Y - \frac{2}{n} \tilde{X}^T \tilde{X} Y + \tilde{X}^T \hat{\Sigma} \tilde{X}$$

$$= \frac{1}{n} \|Y - \tilde{X} \hat{\beta}\|^2 - \hat{\beta}^T \left(\frac{1}{n} \tilde{X}^T \tilde{X} - \hat{\Sigma}\right) \hat{\beta}$$

$$= \frac{1}{n} \|Z \beta + \varepsilon - (\tilde{Z} + \tilde{W}) \hat{\beta}\|^2 - \hat{\beta}^T D \hat{\beta}$$

$$= \frac{1}{n} \|\Delta\|^2 + \frac{1}{n} \varepsilon^T \varepsilon + \frac{2}{n} \varepsilon^T \Delta - \frac{2}{n} \varepsilon^T \tilde{W} \hat{\beta} - \frac{2}{n} \beta^T \tilde{W}^T \Delta + \hat{\beta}^T \left(\frac{1}{n} \tilde{W}^T \tilde{W} - D\right) \hat{\beta}$$

where $\Delta := Z \beta - \hat{Z} \hat{\beta}$. We thus have

$$|\hat{\sigma}^2 - \sigma^2| \leq \frac{1}{n} \varepsilon^T \varepsilon - \sigma^2 + \left\|\frac{1}{n} \tilde{W}^T \tilde{W} - D\right\|_{ap} \|\hat{\beta}\|^2 + \frac{2}{n} |\varepsilon^T \tilde{W} \hat{\beta}|$$

$$+ \frac{1}{n} \|\Delta\|^2 + \frac{2}{n} |\varepsilon^T \Delta| + \frac{2}{n} |\beta^T \tilde{W}^T \Delta|.$$  

We upper bound the first three terms on the RHS. Note that the rate of $\|\hat{\beta} - \beta\|$ in Theorem 4 implies

$$\mathbb{P}\left\{|\|\hat{\beta}\| \leq 1 \lor \|\beta\|\right\} \geq 1 - (p \lor n)^{-c}.$$  

(150)

From Lemma 15 and Lemma 18 with the same discretization arguments in the proof of Lemma 21, we obtain

$$\frac{1}{n} \varepsilon^T \varepsilon - \sigma^2 \lesssim \sqrt{\log(p \lor n) \frac{n}{p}}, \quad \left\|\frac{1}{n} \tilde{W}^T \tilde{W} - D\right\|_{ap} \lesssim \sqrt{K \log(p \lor n) \frac{n}{nm^2}},$$  

(151)
with probability greater than \(1 - (p \lor n)^{-c}\). For the third term, using the analogous arguments of proving Lemma 17 by replacing \(Z_k\) by \(\varepsilon\), we can derive

\[
P \left\{ \frac{1}{n} |\hat{\beta}^T \tilde{W} v| \leq \left( \frac{\|v\|}{\sqrt{m}} + \rho \|v\|_1 \right) \sqrt{\frac{K \log (p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c}
\]

for any fixed \(v \in \mathbb{R}^K\). By choosing \(v = \beta\) and \(v = e_k\) for \(k \in [K]\) and noting that

\[
\frac{1}{n} |\hat{\beta}^T \tilde{W} \beta| \leq \frac{1}{n} |\varepsilon^T \tilde{W} \beta| + \frac{1}{n} \|\varepsilon^T \tilde{W}\|_\infty \sqrt{K} \|\beta - \beta\|,
\]

we have

\[
P \left\{ \frac{1}{n} |\hat{\beta}^T \tilde{W} \beta| \leq C_{\min}^{-1} \left( 1 \lor \frac{\|\beta\|}{\sqrt{m}} \right) \sqrt{\frac{K \log (p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c}.
\]

(152)

We proceed to bound the remaining three terms by studying \(\Delta\) first. From (133), recall that

\[\Delta = Z(\beta - \hat{\beta}) + Z \Delta \hat{\beta}.\]

Since

\[
\frac{1}{n} \|\Delta\|^2 \leq \frac{1}{n} \|Z(\hat{\beta} - \beta)\|^2 + \frac{1}{n} \|Z^T Z\|_\infty \|\Delta \hat{\beta}\|_1 \|\hat{\beta}\|_1,
\]

\[
\frac{1}{n} |\tilde{W} \Delta| \leq \frac{1}{n} |\varepsilon^T \hat{\beta} - \beta| + \frac{1}{n} \|\varepsilon^T Z\|_\infty \|\Delta \hat{\beta}\|_1 \|\hat{\beta}\|_1,
\]

using (135), (38), \(\|\hat{\beta}\|_1 \leq \sqrt{K} \|\beta\|\), (150), \(K \log (p \lor n) = O(n)\) and invoking the event \(\mathcal{E}\), Lemma 15 and Assumption 3 yield

\[
P \left\{ \frac{1}{n} \|\Delta\|^2 + \frac{2}{n} |\tilde{W} \Delta| \leq C_{\min}^{-1} \left( 1 \lor \frac{\|\beta\|}{m} \right) \sqrt{\frac{K \log (p \lor n)}{n}} \right\} \geq 1 - (p \lor n)^{-c}.
\]

(153)

Finally, noting that

\[
\frac{1}{n} |\tilde{W} \Delta|^2 \\
\leq \frac{1}{n} |\hat{\beta}^T \tilde{W}^T \Delta| + \frac{1}{n} |(\hat{\beta} - \beta)^T \tilde{W}^T \Delta| \\
\leq \frac{1}{n} \|\hat{\beta}^T \tilde{W}^T Z\|_\infty \sqrt{K} \|\beta - \beta\| + \frac{1}{n} \|\hat{\beta}^T \tilde{W}^T Z\|_\infty \|\Delta \hat{\beta}\|_1 \|\hat{\beta}\|_1 \\
+ \frac{1}{n} |(\hat{\beta} - \beta)^T \tilde{W}^T Z(\hat{\beta} - \beta)| + \frac{1}{n} \|\tilde{W}^T Z\|_\infty \sqrt{K} \|\beta - \beta\| \|\Delta \hat{\beta}\|_1 \|\hat{\beta}\|_1,
\]

using the same arguments with Lemma 17 bounds from above the terms on the RHS except \(n^{-1} |(\hat{\beta} - \beta)^T \tilde{W}^T Z(\hat{\beta} - \beta)|\). We control this remaining term by noting that

\[
\frac{1}{n} |(\hat{\beta} - \beta)^T Z^T \tilde{W} (\hat{\beta} - \beta)| \leq \|\hat{\beta} - \beta\|^2 \sup_{v \in \mathcal{S}^{K-1}} \frac{1}{n} |v^T Z^T \tilde{W} v|.
\]

(154)

The sup term is handled by display (146). Therefore, collecting (151) – (154) gives the desired result. \(\Box\)
B.16 Proof of lemmas for proving Theorem 8

B.16.1 Proof of Lemma 26

The result for \(\lambda_K(A_T^T A_I)\) follows by using the definition of \(I\) and observing that
\[
\lambda_K(A_T^T A_I) = \inf_{v \in S^{K-1}} \|A_I v\|^2 \\
= \inf_{v \in S^{K-1}} \sum_{k=1}^K \sum_{i \in I_k} (A^T_{ki} v)^2 \\
= \inf_{v \in S^{K-1}} \sum_{k=1}^K |I_k| v_k^2 = \min_k |I_k|.
\]

The proof of \(|A_I M_I^{-1}|_{op}^2\) follows from
\[
|A_I M_I^{-1}|_{op}^2 = |M_I^{-1} A_T^T A_I M_I^{-1}|_{op} = |M_I^{-1}|_{op} = \lambda_K^{-1}(A_T^T A_I).
\]

Similarly, \((ii)\) follows from the fact that
\[
|AM^{-1}|_{op}^2 = |M^{-1} A^T A M^{-1}|_{op} = |M^{-1}|_{op} = \Lambda_\min^{-1}
\]

and
\[
|AM^{-1/2}|_{op}^2 = |M^{-1/2} A^T A M^{-1/2}|_{op} = 1.
\]

\[\square\]

B.16.2 Proof of Lemma 27

By Weyl’s inequality, we have
\[
\frac{1}{n} \lambda_{\min}(Z^T Z) \geq C_{\min} - \|\Sigma^z - n^{-1} Z^T Z\|_{op}
\]
\[
\frac{1}{n} \lambda_{\max}(Z^T Z) \leq C_{\max} + \|\Sigma^z - n^{-1} Z^T Z\|_{op}.
\]

Under Assumption 4 and \(K \log(p \vee n) = O(n)\), invoking Lemma 21 concludes the proof of \((i)\).

We proceed to prove \((ii)\) by letting \(U_i := M^{-1} A^T X_i\), for all \(1 \leq i \leq n\) and noting that
\[
\text{Cov}(U_i) = M^{-1} A^T (\Sigma^z A^T + \Sigma^w)AM^{-1} = \Sigma^z + M^{-1} A^T \Sigma^w A M^{-1}.
\]

Thus, invoking (Vershynin, 2012, Theorem 5.39, Remark 5.40) yields
\[
P \left\{ \left\| \frac{1}{n} U^T U - (\Sigma^z + M^{-1} A^T \Sigma^w A M^{-1}) \right\|_{op} \lesssim C_{\max} \sqrt{\frac{K \log(p \vee n)}{n}} \right\} \geq 1 - p^{-K},
\]
where we also use \(|\Sigma^z + M^{-1} A^T \Sigma^w A M^{-1}|_{op} \leq C_{\max} + C_W \Lambda_{\min}^{-1} \lesssim C_{\max}\) from Lemma 26. Invoking \(K \log(p \vee n) = O(n)\) and Weyl’s inequality concludes the proof of \((ii)\).

To prove \((iii)\), from (Vershynin, 2012, Theorem 5.39, Remark 5.40), we have, with probability greater than \(1 - p^{-|J|}\),
\[
\left\| \frac{1}{n} W_J^T W_J - \Sigma_{J,J}^w \right\|_{op} \lesssim C_W \left( \sqrt{\frac{|J| \log(p \vee n)}{n}} + \frac{|J| \log(p \vee n)}{n} \right).
\]

With the same probability, applying Weyl’s inequality gives \(\lambda_{\max}(W_J^T W_J) \lesssim nC_{w,J}\) as desired. Finally, the proof of \((iv)\) follows from the same arguments of \((ii)\).
B.16.3 Proof of Lemma 28

To prove (i), note that \( \| \hat{M}^{-1} \|_{op} = \lambda_{\min}^{-1}(\hat{M}) \) and

\[
\hat{M} = M + \hat{M} - M = M^{1/2} \left[ I_K + M^{-1/2}(\hat{M} - M)M^{-1/2} \right] M^{1/2}.
\]

It follows from an application of Weyl’s inequality that

\[
\lambda_{\min}(\hat{M}) \geq \lambda_{\min}(M) \Lambda_{\min} \left( I_K + M^{-1/2}(\hat{M} - M)M^{-1/2} \right)
\geq \lambda_{\min}(M) \left[ 1 - \| M^{-1/2}(\hat{M} - M)M^{-1/2} \|_{op} \right].
\]

It then suffices to upper bound the operator norm. Note that

\[
\| M^{-1/2}(\hat{M} - M)M^{-1/2} \|_{op}
\leq \| M^{-1/2}(\hat{A} - A)T(\hat{A} - A)M^{-1/2} \|_{op} + 2\| M^{-1/2}A^T(\hat{A} - A)M^{-1/2} \|_{op}
\leq \| \hat{A} - A \|_{op} \| M^{-1} \|_{op} + 2\| \hat{A} - A \|_{op} \| AM^{-1/2} \|_{op} \| M^{-1/2} \|_{op}.
\]

On the event \( E_f \) defined by (105), invoking (ii) of Lemma 26 yields

\[
\| M^{-1/2}(\hat{M} - M)M^{-1/2} \|_{op} \lesssim R_M^2 + R_M \overset{(106)}{=} o(1).
\]

Thus, we conclude \( \| \hat{M}^{-1} \|_{op} \lesssim \Lambda_{\min}^{-1} \). The second statement of (i) follows from

\[
\| \hat{A} \hat{M}^{-1} \|_{op}^2 = \| \hat{M}^{-1} \|_{op} \lesssim \Lambda_{\min}^{-1}.
\]

To prove (ii), we have

\[
\| W A \hat{M}^{-1} - W A M^{-1} \|_{op}^2 \leq 2\| W(\hat{A} - A) \hat{M}^{-1} \|_{op}^2 + 2\| W A(\hat{M}^{-1} - M^{-1}) \|_{op}^2
\]

On the event \( E_f \), using \( \hat{A}_I = A_I \) and part (i) gives

\[
\| W(\hat{A} - A) \hat{M}^{-1} \|_{op}^2 \leq \| W \|_{op}^2 \| \hat{A}_I - A_I \|_{op}^2 \| \hat{M}^{-1} \|_{op}^2 \lesssim nC_{wJ} \Lambda_{\min}^{-1} R_M^2,
\]

(157)

On the other hand, we have

\[
\| W A(\hat{M}^{-1} - M^{-1}) \|_{op}^2
= \| W A M^{-1}(\hat{M} - M)\hat{M}^{-1} \|_{op}^2
\leq 2\| W A M^{-1}(\hat{A} - A) \hat{A} \hat{M}^{-1} \|_{op}^2 + 2\| W A M^{-1}A^T(\hat{A} - A)\hat{M}^{-1} \|_{op}^2
\leq 2\| W A M^{-1/2} \|_{op}^2 \Lambda_{\min}^{-1} R_M^2 \| \hat{A} \hat{M}^{-1} \|_{op}^2
+ 2\| W A M^{-1/2} \|_{op}^2 \| AM^{-1/2} \|_{op}^2 \| R_A \| \hat{M}^{-1} \|_{op}^2.
\]

Using the fact that \( \| W A M^{-1/2} \|_{op}^2 \lesssim nC_W \) on the event \( E_f \), together with (i) above and (ii) of Lemma 26, we conclude

\[
\| W A(\hat{M}^{-1} - M^{-1}) \|_{op}^2 \lesssim n\Lambda_{\min}^{-1} R_M^2.
\]

(158)
The proof of (ii) is finished by combining (157) and (158) and noting that condition (106) implies
\[ \frac{|J|}{\Lambda_{\min}} \log(p \vee n) \leq \frac{\|A_J\|_0 \log(p \vee n)}{\Lambda_{\min} n} = o(1). \]

Finally, the proof of (iii) follows by part (ii), invoking \( E_f \) and noting that
\[ \|X \hat{A} M^{-1} - X A M^{-1}\|_{op}^2 \]
\[ \leq 2\|Z(A - \hat{A})^T \hat{A} M^{-1}\|_{op}^2 + 2\|W \hat{A} M^{-1} - W A M^{-1}\|_{op}^2 \]
\[ \lesssim nC_{\max} R_M^2. \]

\[ \square \]

C Auxiliary results

C.1 Convergence rate of \( \hat{\beta}_{\text{full}} \) defined in (24)

Proposition 29. Under the same conditions of Theorem 4, assume \( s_J \log(p \vee n) = o(n \Lambda_{\min}) \) holds with \( s_J := \|A_J\|_0 \) and \( \Lambda_{\min} := \lambda_K(A^T A) \). Then, with probability greater than \( 1 - (p \vee n)^{-c} \) for some constant \( c > 0 \), \( \hat{\beta}_{\text{full}} \) in (24) satisfies
\[ \min_{P \in H_K} \|\hat{\beta}_{\text{full}} - P \beta\| \]
\[ \lesssim \left( 1 \vee \frac{\|\beta\|}{\sqrt{m}} \right) \sqrt{\frac{K \log(p \vee n)}{n}} + (1 \vee \|\beta\|) \sqrt{\left( \frac{(p \vee s_J K) \log(p \vee n)}{\Lambda_{\min} n} \right)}. \]

Proof. We only give a brief sketch of the proof. Suppose \( P = I_K \). From the expression of (24) and using model (1), one can easily derive
\[ \hat{\Sigma}^z(\hat{\beta}_{\text{full}} - \beta) = \left( \frac{1}{n} Z^T Z - \hat{\Sigma}^z \right) \beta + \frac{1}{n} Z^T \varepsilon \]
\[ + (\hat{A}^T \hat{A})^{-1} \hat{A}^T \left[ (A - \hat{A}) \frac{1}{n} Z^T Y + \frac{1}{n} W^T Y \right]. \]

The \( \ell_2 \)-norm of the first term can be upper bounded by a slight modification of Lemma 21 and the analysis of the second term follows immediately from Lemmas 13 and 15. Finally, we can upper bound the \( \ell_2 \) norm of the last term by
\[ \| (\hat{A}^T \hat{A})^{-1/2} \|_{op} \left[ \|A - \hat{A}\|_{op} \frac{\sqrt{K}}{n} \|Z^T Y\|_{\infty} + \frac{\sqrt{p}}{n} \|W^T Y\|_{\infty} \right]. \]

The result then can be derived by invoking Lemmas 15, 25 and 28. \[ \square \]
Continuing the same reasoning as before, we conclude that with the difference that we now choose \( M \) arguments as in the proof of Theorem 8 and use the fact that \( Z \). This display suggests to predict \( Y \) and predict \( Y \) by

\[
\tilde{Y}_I = \tilde{Z}_I^T \beta = X_I^T A_I \left[ \text{Cov}(A_I^T X_I) \right]^{-1} \text{Cov}(A_I^T X_I, Y).
\]

This display suggests to predict \( Y \) from the data \((X_i, Y_i)_{i=1}^n\) by

\[
\hat{Y}_I = X_I^T \tilde{A}_I \left( \tilde{A}_I^T X_I^T X_I \tilde{A}_I \right)^{-1} \tilde{A}_I^T X_I^T Y := X_I^T \hat{\theta}_I
\]

where \( \tilde{A}_I \) is obtained from (54) – (55) and \( M^+ \) denotes the Moore-Penrose inverse of any matrix \( M \). The following theorem establishes the in-sample prediction risk \( E\|Z \beta - \hat{Y}_I\|^2 \) and new data prediction risk \( E\|Z \beta - \hat{Y}_*(I)\|^2 \) where \( \hat{Y}_*(I) = X_I^T \hat{\theta}_I \) for a new data point \( X_* \).

**Theorem 30.** Let model (1) and Assumptions 1 – 4 hold. Further assume \( J_1 = \emptyset \). One has

\[
\frac{1}{n} \mathbb{E} \left[ \|Z \beta - \hat{Y}_I\|^2 \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2}{m} C_W.
\]

**Proof.** We write \( \tilde{Z} := X_I^T \tilde{A}_I \) and \( \beta^{LS} := (\tilde{Z}^T \tilde{Z})^+ \tilde{Z}^T Y \) such that \( \hat{Y}_I = \tilde{Z} \beta^{LS} \). Also let \( \tilde{P} = \tilde{Z} (\tilde{Z}^T \tilde{Z})^+ \tilde{Z}^T \) and \( \tilde{P}^\perp = I_n - \tilde{P} \). Consider the event \( \mathcal{E}_f \) defined in (105). We follow the same arguments as in the proof of Theorem 8 and use the fact that \( \tilde{A}_I = A_I \) on \( \mathcal{E}_f \) together with \( J_1 = \emptyset \), with the difference that we now choose \( M_I = A_I^T A_I \) such that

\[
\|\tilde{P}^\perp Z \beta\| = \|\tilde{P}^\perp (Z - \tilde{Z} M_I^{-1}) \beta\|
\]

\[
= \|\tilde{P}^\perp (Z - Z A_I^T A_I M_I^{-1} - W_I A_I M_I^{-1}) \beta\|
\]

\[
\leq \|W_I A_I (A_I^T A_I)^{-1} \beta\|^2.
\]

Continuing the same reasoning as before, we conclude that

\[
\frac{1}{n} \mathbb{E} \left[ \|Z \beta - \hat{Y}_I\|^2 1_{\mathcal{E}_f} \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2}{m} C_W,
\]

which finishes the proof.

**C.3 An auxiliary lemma**

**Lemma 31.** Let \( \tilde{Z}_I \) and \( \tilde{Z} \) be defined in (159) and (50), respectively. Then

\[
\mathbb{E} \left[ \beta^T Z - \beta^T \tilde{Z}_I \right]^2 = \beta^T \left[ \Omega + A_I^T A_I (A_I^T \Sigma_I^w A_I)^{-1} A_I^T A_I \right]^{-1} \beta \leq \frac{C_W \|\beta\|^2}{\lambda_{\min}(A_I^T A_I)},
\]

\[
\mathbb{E} \left[ \beta^T Z - \beta^T \tilde{Z} \right]^2 = \beta^T \left[ \Omega + A^T A (A^T \Sigma^w A)^{-1} A^T A \right]^{-1} \beta \leq \frac{C_W \|\beta\|^2}{\lambda_{\min}(A^T A)}.
\]
Proof. Let \( \bar{\Theta}(I) = (\Sigma^w + \Pi_I^T \Sigma_I^w \Pi_I)^{-1} \Sigma^w \) as (159). Using \( \tilde{X}_I = Z + \tilde{W}_I \) and the independence of \( Z \) and \( W \) give

\[
\mathbb{E} \left[ \beta^T Z - \beta^T \bar{Z}(I) \right]^2 = \mathbb{E} \left[ \beta^T Z - \beta^T \bar{\Theta}(I)^T Z \right]^2 + \mathbb{E} \left[ \beta^T \bar{\Theta}(I) \tilde{W}_I \right]^2 \\
= \beta^T (I_K - \bar{\Theta}(I))^T \Sigma^w (I_K - \bar{\Theta}(I)) \beta + \beta^T \bar{\Theta}(I)^T \Pi_I^T \Sigma_I^w \Pi_I \bar{\Theta}(I) \beta.
\]

Plugging in the expression of \( \bar{\Theta}(I) \) and simplifying (here applying the Woodbury matrix identity to \( \bar{\Theta}(I) \)), we find

\[
\mathbb{E} \left[ \beta^T Z - \beta^T \bar{Z}(I) \right]^2 = \beta^T \left[ \Omega + A_I^T A_I (A_I^T \Sigma_I^w A_I)^{-1} A_I^T A_I \right]^{-1} \beta,
\]

which is upper bounded by

\[
\frac{\|\beta\|^2}{\lambda_{\min}(A_I^T A_I)} \frac{1}{\lambda_{\min}(A_I (A_I^T \Sigma_I^w A_I)^{-1} A_I^T)} \leq C_W \frac{\|\beta\|^2}{\lambda_{\min}(A_I^T A_I)}
\]

The same arguments can be used to prove the second display. \( \square \)

C.4 Proof of lemma 9 and lemma 10

C.4.1 Proof of Lemma 9

We write the eigen-decomposition of \( \Sigma \) as

\[
\Sigma = U D U^T = \sum_{k=1}^K d_k u_k u_k^T + \sum_{k>K} d_k u_k u_k^T := U_K D_K U_K^T + \sum_{k>K} d_k u_k u_k^T
\]

with \( d_1 \geq d_2 \geq \cdots \geq d_p \). When \( \Sigma^w = \tau^2 I_p \), we have \( d_k = \tau^2 \) for \( k \geq K + 1 \). Thus, we obtain

\[
B \Sigma^w B^T = U (D - \tau^2 I_p) U^T = U_K (D_K - \tau^2 I_K) U_K^T
\]

such that \( U_K = B(\Sigma^w)^{1/2} Q (D_K - \tau^2 I_K)^{-1/2} \) for any orthogonal matrix \( Q \), the result of part (1) then follows.

We proceed to prove (2) and write \( B \Sigma^w B^T = V_K \text{diag}(\lambda) V_K^T = \sum_{k=1}^K \lambda_k v_k v_k^T \) with \( \lambda_1 > \lambda_2 > \cdots > \lambda_K \) that satisfy condition (46). We first prove the statement of \( \|U_K - B Q\| \). By the sin(\( \theta \)) theorem (Davis and Kahan, 1970; Yu et al., 2014), we can choose the signs of \( V_K = (v_1, \ldots, v_K) \) such that, for any \( 1 \leq k \leq K \),

\[
\|u_k - v_k\| \leq \frac{\sqrt{2} \|\Sigma^w\|_{op}}{\min(|d_{k-1} - \lambda_k|, |\lambda_k - d_{k+1}|)}.
\]

with \( d_0 := \infty \). On one hand, by the reverse triangle inequality followed by Weyl’s inequality we have

\[
|d_{k-1} - \lambda_k| \geq |\lambda_{k-1} - \lambda_k| - |d_{k-1} - \lambda_{k-1}|
\]

(46)

\[
\geq c_* \lambda_K - \|\Sigma^w\|_{op}
\]

\[
= (c_* \xi - 1) \|\Sigma^w\|_{op}
\]

\[
\geq \xi \|\Sigma^w\|_{op},
\]

77
where we recall that \( \xi = \lambda_K(B\Sigma^2B^T)/\|\Sigma^w\|_{op} \) and in the last two steps use \( c_\ast, \xi > 1 \). On the other hand, a similar argument yields

\[
|\lambda_k - d_{k+1}| \geq |\lambda_k - \lambda_{k+1}| - |d_{k+1} - \lambda_{k+1}| \gtrsim \xi\|\Sigma^w\|_{op}
\]

for \( 1 \leq k \leq K - 1 \). For \( k = K \), we have

\[
|\lambda_k - d_{k+1}| \geq \lambda_K - \|\Sigma^w\|_{op} \gtrsim \xi\|\Sigma^w\|_{op}.
\]

We thus conclude \( \|u_k - v_k\| = O(\xi^{-1}) \) hence \( \|U_K - V_K\| = O(\sqrt{K}/\xi) \). Since there exists an invertible matrix \( Q \) such that \( V_K = BQ \), we conclude that \( \|U_K - BQ\| = O(\sqrt{K}/\xi) \).

Finally, we upper bound \( \mathbb{E}[Y_{PCR} - Y_{FR}]^2 \) by first recalling that

\[
Y_{PCR}^* = X^TU_K(U_K^T\Sigma U_K)^{-1}U_K^T\beta, \quad Y_{FR}^* = X^TV_K(V_K^T\Sigma V_K)^{-1}V_K^T\beta
\]

from \( B = V_KQ \) and the fact that \( Q \) is invertible. Letting \( \bar{\beta} = B\Sigma^z\beta \), we then have

\[
\mathbb{E}[Y_{PCR}^* - Y_{FR}^*]^2 = \left\| \Sigma^{1/2} [U_K(U_K^T\Sigma U_K)^{-1}U_K^T - V_K(V_K^T\Sigma V_K)^{-1}V_K^T] \bar{\beta} \right\|^2
\]

where in the third line we use

\[
\Sigma^{1/2}U_K(U_K^T\Sigma U_K)^{-1}U_K^T = UD^{1/2}U^TU_KD_K^{-1}U_K^TU^2U^T = U_KU_K^T
\]

and define \( \bar{U}_K = \Sigma^{1/2}V_K(V_K^T\Sigma V_K)^{-1/2} \). Recalling that \( \bar{\beta} = B\Sigma^z\beta, \\|\Sigma^{-1/2}\bar{\beta}\|_{op} \) can be upper bounded by

\[
\|\Sigma^{-1/2}\bar{\beta}\|^2 \leq (\beta^T\Sigma^z\beta)\|\Sigma^{1/2}B\Sigma^2B^T\Sigma^{-1/2}\|_{op}
\leq (\beta^T\Sigma^z\beta)\|\Sigma^{-1/2}B\Sigma^2\Sigma B^T\Sigma^{-1/2}\|_{op}
\leq (\beta^T\Sigma^z\beta)(1 + \|\Sigma^w\|_{op}/d_K) \quad \text{(by } \Sigma = B\Sigma^2B^T + \Sigma^w\)
\leq \beta^T\Sigma^z\beta
\]

(161)

where the last line uses \( \xi \to \infty \) and

\[
d_K \geq \lambda_K - \|\Sigma^w\|_{op} \gtrsim \xi\|\Sigma^w\|_{op}
\]

(162)

from Weyl’s inequality. What remains is to bound \( \|U_KU_K^T - \bar{U}_K\bar{U}_K^T\|^2 \). Recall that the columns of \( U_K \) and \( \bar{U}_K \) are, respectively, the first \( K \) eigenvectors of \( \Sigma^{1/2}U_KU_K^T\Sigma^{1/2} \) and \( \Sigma^{1/2}V_KV_K^T\Sigma^{1/2} \). Without loss of generality, we can assume \( u_k^T\bar{u}_k \geq 0 \) for \( 1 \leq k \leq K \) since \( \|U_KU_K^T - \bar{U}_K\bar{U}_K^T\|^2 \) is invariant to orthogonal transformations of the columns of \( \bar{U}_K \). Thus, an application of \( \sin(\theta) \) Theorem yields

\[
\|u_k - \bar{u}_k\| \leq \sqrt{2}\|\Sigma^{1/2}(U_KU_K^T - V_KV_K^T)\Sigma^{1/2}\|_{op} \min\left(\frac{d_k}{|d_{k-1} - d_k|, |d_k - d_{k+1}|}\right), \quad 1 \leq k \leq K
\]

78
where \( \tilde{d}_k \) for \( 1 \leq k \leq K \) are the eigenvalues of \( \Sigma^{1/2}V_KV_K^T\Sigma^{1/2} \) and we denote \( \tilde{d}_{K+1} = 0 \). The numerator is bounded by

\[
2\sqrt{2}\|\Sigma\|_{op}\|U_K - V_K\|_{op} \leq 2\sqrt{2}d_1\|U_K - V_K\|.
\]

Since \( d_1 \leq \lambda_1 + \|\Sigma^w\|_{op} \leq \lambda_1 \) from Weyl’s inequality, the previous result of \( \|U_K - V_K\| \) yields

\[
\|\tilde{A}\|_{op} := \|\Sigma^{1/2}(U_KU_K^T - V_KV_K^T)\Sigma^{1/2}\|_{op} \lesssim \sqrt{K}\lambda_1/\xi = \kappa\sqrt{K}\|\Sigma^w\|_{op},
\]

with \( \kappa := \lambda_1/\lambda_K \). By similar arguments as before, we can lower bound the denominator as

\[
|\tilde{d}_{k-1} - d_k| \geq |d_{k-1} - d_k| - |\tilde{d}_{k-1} - d_{k-1}|
\]

\[
\geq |\lambda_{k-1} - \lambda_k| - 2\|\Sigma^w\|_{op} - \|\tilde{A}\|_{op} \gtrsim \xi\|\Sigma^w\|_{op},
\]

for \( 1 \leq k \leq K \), and

\[
|d_k - \tilde{d}_{k+1}| \geq |d_k - d_{k+1}| - |\tilde{d}_{k+1} - d_{k+1}| \gtrsim \xi\|\Sigma^w\|_{op}
\]

for \( 1 \leq k \leq K - 1 \) together with \( |d_K - \tilde{d}_{K+1}| = d_K \gtrsim \xi\|\Sigma^w\|_{op} \) from (162). Combining these three lower bounds with (163) concludes \( \|u_k - \tilde{u}_k\| = O(\kappa\sqrt{K}/\xi) \) hence \( \|U_k - \tilde{U}_k\| = O(K\kappa/\xi) \). In conjunction with (161) and \( \|U_KU_K^T - \tilde{U}_K\tilde{U}_K^T\| \leq \|U_K - \tilde{U}_K\|^2 \), we obtain

\[
E\left[|Y^*_{PCR} - Y^*_{FR}|^2 \lesssim (\beta^T\Sigma^z\beta)(K\kappa/\xi)^2 \right].
\]

This completes the proof. \qed

### C.4.2 Proof of Lemma 10

We work on the event on which the inverses in the formula for \( \tilde{Z}_{(i)} \) exist, so \( \tilde{Z}_{(i)} \) is well defined. By writing \( \tilde{X} = X_\hat{I}_\hat{A}_\hat{T}(\hat{A}_\hat{T}\hat{A}_\hat{T})^{-1} \), the least squares solution satisfies

\[
\tilde{\beta} = \left( \tilde{Z}_{(i)}\tilde{I}_{(i)} \right)^{-1} \tilde{Z}_{(i)} \tilde{Y} = (\tilde{\Sigma}^z)^{-1} \left( \tilde{\Sigma}^z + \tilde{\Pi}_\hat{I}_\hat{T}\tilde{\Sigma}^w\tilde{\Pi}_\hat{I} \right) \left( \tilde{X}^T\tilde{X} \right)^{-1} \tilde{X}^T\tilde{Y}.
\]

Recall from (14) that \( \tilde{\beta} = (\tilde{\Sigma}^z)^{-1}n^{-1}\tilde{X}^T\tilde{Y} \). We only need to show

\[
\tilde{\Sigma}^z + \tilde{\Pi}_\hat{I}_\hat{T}\tilde{\Sigma}^w\tilde{\Pi}_\hat{I} = \frac{1}{n}\tilde{X}^T\tilde{X}.
\]

Since our estimation of \( \tilde{\Sigma}^z \) gives

\[
\frac{1}{n}\tilde{X}^T\tilde{X} = \tilde{\Sigma}^z + D
\]

where \( D = \text{diag}(d_1, \ldots, d_K) \) with \( d_k \) defined in (74) for \( k \in [\hat{K}] \). It then suffices to show \( D = \tilde{\Pi}_\hat{I}_\hat{T}\tilde{\Sigma}^w\tilde{\Pi}_\hat{I} \). Note that \( \tilde{\Pi}_\hat{I}_\hat{T}\tilde{\Sigma}^w\tilde{\Pi}_\hat{I} \) is also diagonal. We conclude the proof by computing

\[
\left[ \tilde{\Pi}_\hat{I}_\hat{T}\tilde{\Sigma}^w\tilde{\Pi}_\hat{I} \right]_{kk} = \frac{1}{|\hat{I}_k|^2} \sum_{i \in \hat{I}_k} \tilde{\Sigma}^w_{ii} \overset{\text{(27)}}{=} \frac{1}{|\hat{I}_k|^2} \sum_{i \in \hat{I}_k} \left[ \frac{1}{n}X_i^TX_i - \hat{\Sigma}^z_{kk} \right] \overset{\text{(53)}}{=} d_k
\]