Adaptive Estimation of Multivariate Regression with Hidden Variables

Xin Bing∗ Yang Ning† Yaosheng Xu‡

Abstract

A prominent concern of scientific investigators is the presence of unobserved hidden variables in association analysis. Ignoring hidden variables in the analysis often yields biased statistical results and misleading scientific conclusions. Motivated by this practical issue, this paper studies the estimation of the coefficient matrix $\Theta^*$ in multivariate regression with hidden variables, 

$$ Y = (\Theta^*)^T X + (B^*)^T Z + E, $$

where $Y$ is a $m$-dimensional response vector, $X$ is a $p$-dimensional vector of observable features, $Z$ represents a $K$-dimensional vector of unobserved hidden variables, possibly correlated with $X$, and $E$ is an independent error. The number of hidden variables $K$ is unknown and both $m$ and $p$ are allowed but not required to grow with the sample size $n$.

Since only $Y$ and $X$ are observable, we first provide necessary conditions for the identifiability of $\Theta^*$. The same set of conditions are shown to be sufficient when the error $E$ is homoscedastic. Our identifiability proof is constructive and leads to a novel and computationally efficient estimation algorithm for $\Theta^*$, called HIVE. The first step of the algorithm is to estimate the best linear prediction of $Y$ given $X$, in which the unknown coefficient matrix exhibits an additive decomposition of $\Theta^*$ and a dense matrix originated from the correlation between $X$ and the hidden variable $Z$. Under the row sparsity assumption on $\Theta^*$, we propose to minimize a penalized least squares loss by regularizing $\Theta^*$ via a group-lasso penalty and regularizing the dense matrix via a multivariate ridge penalty. Non-asymptotic deviation bounds of the in-sample prediction error are established. Our second step is to estimate the row space of $B^*$ by leveraging the covariance structure of the residual vector from the first step. In the last step, we remove the effect of hidden variable by projecting $Y$ onto the complement of the estimated row space of $B^*$. Non-asymptotic error bounds of our final estimator, which are valid for any $m, p, K$ and $n$, are established. We further show that under mild assumptions the rate of our estimator matches the best possible rate with known $B^*$ and our estimator is adaptive to the unknown sparsity of $\Theta^*$. The model identifiability, parameter estimation and statistical guarantees are further extended to the setting with heteroscedastic errors. Thorough numerical simulations and two real data examples are provided to back up our theoretical results.

Keywords: high-dimensional models, multivariate linear regression, shrinkage, non-sparse estimation, hidden variables, confounding, surrogate variable analysis

∗Department of Statistics and Data Science, Cornell University, Ithaca, NY. E-mail: xb43@cornell.edu.
†Department of Statistics and Data Science, Cornell University, Ithaca, NY. E-mail: yn265@cornell.edu.
‡Department of Statistics and Data Science, Cornell University, Ithaca, NY. E-mail: yx433@cornell.edu.
1 Introduction

Multivariate regression has been widely used to evaluate how predictors are associated with multiple response variables and is ubiquitous in many areas including genomics, epidemiology, social science and economics (Muirhead, 1982; Anderson, 1984; Srivastava and Khatri, 1979; Reinsel and Velu, 1998). Most of the existing research on multivariate regression assumes that the collected predictors are sufficient to explain the responses. However, due to limited resources in practice, oftentimes there still exist unmeasured hidden variables that are associated with the responses. Ignoring the hidden variables often leads to biased estimates.

In this paper, we consider the following multivariate regression with hidden variables. Let $Y \in \mathbb{R}^m$ denote the response vector, $X \in \mathbb{R}^p$ denote the observable predictors and $Z \in \mathbb{R}^K$ be the unobservable hidden variables. The multivariate regression model postulates

$$Y = (\Theta^*)^T X + (B^*)^T Z + E,$$

where $\Theta^* \in \mathbb{R}^{p \times m}$ and $B^* \in \mathbb{R}^{K \times m}$ are unknown deterministic matrices and $E \in \mathbb{R}^m$ is a stochastic error with zero mean and a diagonal covariance matrix $\Sigma_E$. We assume the random error $E$ is independent of $(X, Z)$ and allow the hidden variable $Z$ to correlate with $X$. The number of hidden variables, $K$, is unknown and is typically smaller than $m$. Without loss of generality, we assume $\Sigma = \text{Cov}(X)$ and $\Sigma_Z = \text{Cov}(Z)$ are strictly positive definite and $\text{rank}(B^*) = K$. Otherwise, one might reduce the dimensions of $X$ and $Z$ such that these conditions are met.

Assume that we observe $n$ i.i.d. copies of $(X, Y)$ and stack them together as a design matrix $X \in \mathbb{R}^{n \times p}$ and a response matrix $Y \in \mathbb{R}^{n \times m}$. In practice, the number of response variables $m$ or the number of features $p$ or both of them can be greater than the sample size $n$. The main interest is to identify and estimate $\Theta^*$ so that we can draw valid scientific conclusions on the association between the primary features $X$ and the responses $Y$ in the presence of unobserved hidden variables $Z$.

The proposed model unifies and generalizes the following two strands of research that emerges in a variety of applications.

1. Surrogate variable analysis (SVA) in genomics. The measurements of high-throughput genomic data are often confounded by unobserved factors. To remove the influence of the unobserved confounders, surrogate variable analysis (SVA) based on model (1.1) has been proposed for the analysis of biological data (Leek and Storey, 2007, 2008; Teschendorff et al., 2011; Chakraborty et al., 2012; Gagnon-Bartsch and Speed, 2012; Houseman et al., 2012; Sun et al., 2012). In these applications, the response vector $Y$ is often the gene expression or DNA methylation levels at $m$ sites, which is usually much larger than the sample size $n$. The covariate $X$ is a small set of exposures (e.g., treatment variables), whose dimension $p$ is assumed to be fixed in the theoretical analysis (Lee et al., 2017; Wang et al., 2017; McKennan and Nicolae, 2019). Since $p$ is small, the existing SVA methods apply the ordinary least squares (OLS) $(X^T X)^{-1} X^T Y$ to estimate the main regression effect and then remove the bias of OLS, originated from the correlation between $Z$ and $X$. However, to avoid confounding issues, researchers may collect
as many features as possible and then adjust them in the regression model. In this case, \( p \) can be large and possibly much larger than \( n \), whence the existing SVA methods are not applicable as OLS may not exist. Our work extends the scope of the SVA in the sense that a unified estimation procedure and theoretical justification are developed under model (1.1) where both \( p \) and \( m \) are allowed, but not required, to grow with \( n \). We refer to Section 1.2 for detailed comparisons with existing SVA literature.

2. Structural equation model in causal inference. Model (1.1) can also be framed as linear structural equation models (Hox and Bechger, 1998). Suppose the causal structure among \((X, Z, Y)\) is represented by the directed acyclic graph (DAG) in Figure 1. As shown in this graph, both observed variables \( X \) and hidden variables \( Z \) are the causes of \( Y \), as \((X, Z)\) are the parents of \( Y \). Under the linearity assumption, the causal structure of \((X, Z) \rightarrow Y\) is modeled by equation (1.1). Similarly, the DAG in Figure 1 also implies that \( X \) is the cause of \( Z \), which can be further modeled via

\[
Z = D^T X + W, \tag{1.2}
\]

where \( D \in \mathbb{R}^{p \times K} \) is a deterministic matrix and \( W \in \mathbb{R}^K \) is a random noise independent of \( X \) and \( E \). Since \( Z \) is not observed, model (1.1) and (1.2) can be viewed as linear structural equation models with hidden variables (Diaz, 2017). Using the terminology in causal mediation analysis, the parameter \( \Theta^* \) in (1.1) is interpreted as the direct causal effect of \( X \) on \( Y \), which is often the parameter of interest in the linear structural equations. It is worthwhile to note that the proposed framework is more general than linear structural equation models because model (1.2) is not imposed. In particular, we allow an arbitrary dependence structure between \( X \) and \( Z \), whereas the linear structural equation model assumes \( X \) is a cause of \( Z \) by imposing the independence between \( X \) and \( W \).

Figure 1: Illustration of the DAG under model (1.1) and (1.2)

1.1 Our contributions

We now state our main contributions in this paper.

**Identifiability of \( \Theta^* \).** Since \( Z \) is unobservable, \( \Theta^* \) in model (1.1) is generally not identifiable. Our first contribution is to address a fundamental question of this model that is under what conditions \( \Theta^* \) is identifiable. To motivate our identifiability conditions, we start by rewriting model (1.1). Denote by \((A^*)^T X\) the \( L_2 \) projection of \( Z \) onto the linear space of \( X \) and by
We emphasize that we do not require model (1.2), or equivalently, the independence between $W$ and $X$. For this reason, we use a different notation $A^*$ rather than $D$ to denote the coefficient of the $L_2$ projection. We then decompose the effect of hidden variable $Z$ as $(B^*)^T Z = (A^* B^*)^T X + (B^*)^T W$. Plugging this into (1.1) yields

$$Y = (\Theta^* + A^* B^*)^T X + (B^*)^T W + E := (F^*)^T X + \varepsilon,$$

(1.4)

where $F^* := \Theta^* + L^*$ with $L^* = A^* B^*$ and the new residual vector $\varepsilon := (B^*)^T W + E$ satisfies $E[\varepsilon] = 0$ and $\text{Cov}(X, \varepsilon) = 0$. To disentangle $\Theta^*$ from $L^*$, we impose the following orthogonality restriction between the row spaces of $\Theta^* \in \mathbb{R}^{p \times m}$ and $B^* \in \mathbb{R}^{K \times m}$.

**Assumption 1.** Let $P_{B^*} = B^T (B^* B^T)^{-1} B^* \in \mathbb{R}^{m \times m}$ denote the projection matrix onto the row space of $B^*$. Assume $\Theta^* P_{B^*} = 0$.

In Proposition 1 of Section 2.1, we first show that the above assumption is necessary to identify $\Theta^*$ under model (1.4). In Proposition 2 of Section 2.1, we further show that Assumption 1 is also sufficient when the error $E$ is homoscedastic in the sense that $\Sigma_E = \text{Cov}(E) = \tau^2 I_m$. The covariance structure of the residual vector $\varepsilon$ is crucial to establish this result, which will be detailed in Section 2.1. However, this sufficiency no long holds in the presence of heteroscedastic error, that is, $\Sigma_E$ is a diagonal matrix with unequal entries. Inspired by Zhang et al. (2018), we introduce a mild incoherence condition on the right singular vectors of $B^*$ to identify the row space of $B^*$, which is an important intermediate step towards identifying $\Theta^*$. We show in Proposition 8 of Section 4 that Assumption 1 together with this incoherence condition guarantees the identifiability of $\Theta^*$ under the heteroscedastic case.

Assumption 1 also provides new insights on the interpretation of $\Theta^*$. For an arbitrary $\Theta^*$, we can always write $\Theta^* = \Theta^* P_{B^*} + \Theta^* P_{\perp B^*}$, where $P_{\perp B^*} = I_m - P_{B^*}$. In view of (1.4), $\Theta^* P_{\perp B^*}$ represents the effect of $X$ on $Y$ that cannot be explained by any hidden variables. In the mediation analysis, we can refer to $\Theta^* P_{\perp B^*}$ as “partial” direct effect. Assumption 1 thus imparts these interpretations to the whole $\Theta^*$ by assuming $\Theta^* P_{\perp B^*} = \Theta^*$. While theoretically we can avoid Assumption 1 and establish the identifiability and estimation results for $\Theta^* P_{\perp B^*}$, without loss of generality we impose Assumption 1 to simplify the presentation.

**Estimation of $\Theta^*$.** Our second contribution is to propose a new method for estimating $\Theta^*$. In particular, our approach can handle the case when $p > n$ and the OLS commonly used in the SVA literature does not exist. To deal with the high dimensionality of $\Theta^*$, we assume that there exists a small subset of $X$ that are associated with $Y$ in model (1.1). Such a row-wise sparsity assumption on the coefficient matrix has been widely used in multivariate regression, for instance, Bühlmann and Van de Geer (2011); Bunea et al. (2012); Lounici et al. (2011); Obozinski et al. (2011); Yuan and Lin (2006), just to name a few. Specifically, we assume

$$\Theta^* \in \left\{ \Theta \in \mathbb{R}^{p \times m} : \|\Theta\|_{\ell_0/\ell_2} \leq s_* \right\},$$

(1.5)
where \( s_* \leq p \) and \( \| \Theta \|_{\ell_0/\ell_2} = \sum_{j=1}^p 1_{\{\| \Theta_j \|_2 \neq 0\}} \) is the number of nonzero rows. Our estimation procedure consists of three steps: first estimate the best linear predictor of \( Y \) from \( X \); then estimate the row space of \( B^* \) and finally estimate \( \Theta^* \).

The first step is critical but challenging especially when \( p \) is large. In Section 2.2.1, we propose a new optimization-based approach with a combination of the group-lasso penalty (Yuan and Lin, 2006) and the multivariate ridge penalty. The group-lasso penalty aims to exploit the row-wise sparsity of \( \Theta^* \) in (1.5), while the multivariate ridge penalty regularizes the additional dense signal \( L^* \) due to the hidden variables \( Z \) (see model (1.4)). The proposed procedure is easy to implement and has almost the same complexity as solving a group-lasso problem. We refer to Sections 2.2.1 and 3.4 for detailed discussions of computational and theoretical advantages of our estimator over some competing methods.

Our second step is to estimate the row space of \( B^* \) or equivalently \( P B^* \). When the noise is homoscedastic, we can directly apply the principle component analysis (PCA) to the estimated residual matrix (see Section 2.2.2). The resulting first \( K \) eigenvectors are then used to estimate \( P B^* \). However, PCA may lead to biased estimates under heteroscedastic error, especially when \( m \) is fixed. To deal with heteroscedasticity, we adapt the HeteroPCA algorithm originally proposed by Zhang et al. (2018) to our setting.

In Section 2.2.3, we propose the third step of our procedure to estimate \( \Theta^* \). This step first projects \( Y \) onto the orthogonal complement of the estimated row space of \( B^* \) to remove the effect of hidden variables, and then estimate \( \Theta^* \) by applying the group-lasso to the resulting projected \( Y \).

Our entire procedure is summarized in Algorithm 1, called HIVE, representing HIdden Variable adjustment Estimation. Similarly, the algorithm tailored for the heteroscedastic error is referred to as H-HIVE in Algorithm 3. For the convenience of practitioners, we also provide detailed discussions in Section 5 on practical implementations, including estimation of the number of hidden variables (\( K \)), the consequence of overestimating/underestimating \( K \), the choice of tuning parameters and data standardization.

**Statistical guarantees.** Our third contribution is to establish theoretical properties of our procedure. In Theorem 3 of Section 3, we derive non-asymptotic deviation bounds of the in-sample prediction error, which are valid for any finite \( n, p, m \) and \( K \). The error bounds consist of three parts: a bias term and a variance term from the ridge regularization and an error term from the group-lasso regularization. To understand the advantage of our estimator, we particularize to the orthogonal design and show that our estimator enjoys the optimal rate of group-lasso when there is no hidden variable (i.e., \( L^* = 0 \) in model (1.4)) and it also achieves the optimal rate of the ridge estimator when \( \Theta^* = 0 \). Thus, the rate of our estimator matches the best possible rate even if \( L^* = 0 \) or \( \Theta^* = 0 \) were known a priori.

In Section 3, we provide theoretical guarantees for the estimation of \( \Theta^* \). In particular, we establish in Theorem 5 a general non-asymptotic upper bound of the estimation error of our final estimator \( \hat{\Theta} \) based on any estimator \( \hat{P} \) of \( P B^* \). As expected, the estimation error of \( \hat{\Theta} \) depends on how accurately \( \hat{P} \) estimates \( P B^* \). When \( P B^* \) can be estimated accurately enough,
our estimator \( \tilde{\Theta} \) achieves the optimal rate in the oracle case with known \( B^* \) (see the subsequent paragraph of Theorem 5). However, if the estimation error of \( P_{B^*} \) is relatively large, we can balance this error with the error of the group-lasso to attain a more refined rate via a suitable choice of the regularization parameter. In Theorem 6 of Section 3.2 and Theorem 9 of Section 4, we further establish the non-asymptotic error bounds of our proposed estimators of \( P_{B^*} \) for both homoscedastic and heteroscedastic errors. These results together with Theorem 5 provide the final upper bounds of the estimation error of \( \tilde{\Theta} \). For heteroscedastic errors, we develop a new robust sin \( \Theta \) theorem in Appendix A to control the perturbation of eigenspaces in the Frobenius norm. This theorem is crucial to the proof of Theorem 9 and can be of its own interest.

1.2 Related literature

This work is most related to the literature on surrogate variable analysis (SVA). For model identifiability, Gagnon-Bartsch and Speed (2012); Wang et al. (2017) assumed that there exists a known subset \( J \in \{1, \ldots, m\} \) such that the \( p \times |J| \) submatrix \( \Theta^*_J = 0 \). This set \( J \) is known as “negative control” in the microarray studies. However, this side information is usually unknown in other settings. Another approach by Wang et al. (2017); McKennan and Nicolae (2019) assumes that each row \( \Theta^*_j \in \mathbb{R}^m \) is sparse with \( \|\Theta^*_j\|_0 \leq (m - a)/2 \) for some \( a > K \) and any \( K \times a \) submatrix of \( B^* \) is of rank \( K \). Under this assumption, the sparsity pattern of \( \Theta^* \) differs from (1.5), considered in this work, and this assumption also rules out the possibility that \( B^* \) could be sparse. In the work of Lee et al. (2017), they assumed a similar condition as our Assumption 1. However, when the error is heteroscedastic, Lee et al. (2017) implicitly required \( m \to \infty \) to show the asymptotic identifiability of \( \Theta^* \), see our Remark 6 for more explanations. In contrast, our identifiability result holds for any finite \( n, p, m \) and \( K \). To show the estimation consistency, all existing SVA methods require that \( m \) grows with \( n \) and is typically much larger than \( n \), meanwhile \( p \) is fixed and small, whereas our method provides a more general theoretical framework in which both \( p \) and \( m \) are allowed, but not required, to grow with \( n \).

Chandrasekaran et al. (2012) studied the estimation of Gaussian graphical models with latent variables. In their setting, one can rewrite their estimand as the sum of a low-rank matrix and a sparse matrix (see Hsu et al. (2011); Candès et al. (2011) for other related examples). The regularized maximum likelihood approach is proposed with a combination of the lasso penalty and the nuclear norm penalty. Our problem is related to theirs, because model (1.4) is a regression problem where the coefficient matrix has an additive decomposition of a sparse and a low-rank matrix when \( K \) is much smaller than \( p \) and \( m \). However, our work differs significantly from this strand of research in the following aspects. First, our identifiability Assumption 1 is intrinsically different from theirs. To see this, consider a simple example based on the regression model (1.4) with \( p = m \) and \( K = 1 \). Let \( A^* = e_i \) and \( B^* = e_i^T \), where \( e_i \) is the \( i \)th canonical basis vector of \( \mathbb{R}^p \). The identifiability assumption in Chandrasekaran et al. (2012) does not hold, because the low rank matrix \( L^* = A^*B^* = ee_i^T \) is too sparse and cannot be distinguished from the sparse matrix \( \Theta^* \). However, it is easy to verify \( \Theta^* \in \{ \Theta(I_p - ee_i^T) : \Theta \in \mathbb{R}^{p \times p}\} \) is still identifiable under our Assumption 1 when the error is homoscedastic. One explanation is that the covariance structure of \( \varepsilon = (B^*)^T W + E \) from model (1.4) can assist the identification of
Θ*, whereas this information is ignored if one directly applies the approach in Chandrasekaran et al. (2012). Second, due to the distinct identifiability assumptions, it is not surprising to see that our algorithm (HIVE or H-HIVE) for estimating Θ* is fundamentally different from their regularized maximum-likelihood approach. In particular, our regularized estimation in the first step of our algorithm combines the group-lasso penalty and the ridge penalty. We provide a technical comparison of the ridge penalty and the nuclear norm penalty in Section 3.4.

Recently, Diaz (2017) applied SVA to estimate the causal effect under the structural equation models with hidden variables. As discussed previously, the structural equation models assume (1.2), which is not needed in our modeling framework. Our model (1.4) is derived without imposing any specific model between X and Z. For instance, we allow the true dependence structure between X and Z to be very complicated and highly nonlinear. The estimation method of Diaz (2017) is adapted from the SVA literature, and therefore suffers from the same drawback. In another recent paper, Čevid et al. (2018) proposed a new spectral deconfounding approach to deal with high-dimensional linear regression with hidden confounding variables. In particular, their model can be written as a perturbed linear regression $Y = X^T(\beta + b) + \epsilon$ where $\epsilon \in \mathbb{R}$ is a random noise, $\beta \in \mathbb{R}^p$ is an unknown sparse vector and $b \in \mathbb{R}^p$ is a small perturbation vector. In order to identify $\beta$, they assumed that $\|b\|_2$ is sufficiently close to zero. Unlike this work, we consider a different setting where the response $Y$ is multivariate and, consequently, both our identifiability Assumption 1 and estimation procedures (HIVE and H-HIVE) are completely different from theirs. Our theoretical results in Corollary 7 and its subsequent Remark 4 imply that the convergence rate of our estimator benefits substantially from the multivariate nature of the response, which can be viewed as the blessing of dimensionality.

Outline. In Section 2, we study the identifiability and estimation of Θ* under homoscedastic error. Sufficient and necessary conditions for the identifiability of Θ* are established in Section 2.1. Section 2.2 contains three steps of our estimation procedure. The estimation of $\Theta^* + A^*B^*$ in model (1.4) is stated in Section 2.2.1 and the estimation of the row space of $B^*$ is discussed in Section 2.2.2. The final step of estimating $\Theta^*$ is stated in Section 2.2.3. Section 3.1 is dedicated to the deviation bounds of the in-sample prediction error. The estimation errors of our estimator of $\Theta^*$ together with the errors for estimating the row space of $B^*$ are given in Section 3.2. The extension to heteroscedastic case is studied in Section 4. In Section 5, we discuss several practical considerations, including the selection of $K$, the consequence of overestimating and underestimating $K$, the choice of tuning parameters and data standardization. Simulation results and real data applications are presented in Sections 6 and 7.

1.3 Notation

For any set $S$, we write $|S|$ for its cardinality. For any vector $v \in \mathbb{R}^d$ and some real number $q \geq 0$, we define its $\ell_q$ norm as $\|v\|_q = (\sum_{j=1}^{d} |v_j|^q)^{1/q}$. For any matrix $M \in \mathbb{R}^{d_1 \times d_2}$, $I \subseteq \{1, \ldots, d_1\}$ and $J \subseteq \{1, \ldots, d_2\}$, we write $M_{I,J}$ as the $|I| \times |J|$ submatrix of $M$ with row and column indices corresponding to $I$ and $J$, respectively. In particular, $M_{I}$ denotes the $|I| \times d_2$ submatrix and $M_{J}$ denotes the $d_1 \times |J|$ submatrix. Further write $\|M\|_{\ell_p/\ell_q} = (\sum_{j=1}^{d_2} \|M_{I,J}\|_{\ell_q}^p)^{1/p}$ and denote
by $\|M\|_{\ell_0}$, $\|M\|_{\text{op}}$ and $\|M\|_F$, respectively, the element-wise $\ell_0$ norm, the operator norm and the Frobenius norm of $M$. For any symmetric matrix $M$, we write $\lambda_k(M)$ for its $k$th largest eigenvalue. For any symmetric matrix $M$, we write $\lambda_k(M)$ for its $k$th largest eigenvalue. For any two sequences $a_n$ and $b_n$, we write $a_n \lesssim b_n$ if there exists some positive constant $C$ such that $a_n \leq C b_n$. Both $a_n \asymp b_n$ and $a_n = \Omega(b_n)$ stand for $a_n = O(b_n)$ and $b_n = O(a_n)$. Denote $a \lor b = \max(a, b)$ and $a \land b = \min(a, b)$. Throughout the paper, we will write $\hat{\Sigma} = n^{-1} X^T X$ with non-zero eigenvalues $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$ and $q := \text{rank}(X)$.

2 Identifiability and estimation under homoscedastic noise

As seen in the Introduction, the identifiability of $\Theta^*$ under model (1.1) or equivalently (1.4) needs to be carefully studied due to the presence of hidden variables. We first state necessary conditions for identifying $\Theta^*$ and then show that these conditions are also sufficient when the error $E$ is homoscedastic. Our identifiability procedure is constructive and is further used for estimation.

2.1 Identifiability

In the following proposition, we first establish the necessity of Assumption 1 for identifying $\Theta^*$ under model (1.1), or equivalently, model (1.4). We further show that both Assumption 1 and $\text{rank}(\Sigma_W) = K$ are necessary for identifying $\Theta^*$ if $E[W|X] = 0$ holds, where $\Sigma_W := \text{Cov}(W)$ and $W = Z - (A^*)^T X$ with $A^*$ defined in (1.3).

Proposition 1. Under model (1.4), suppose $Z \in \mathbb{R}^K$ has continuous support and $A^* \neq 0$. Then

(1) Assumption 1 is necessary for identifying $\Theta^*$.

(2) If additionally $E[W|X] = 0$ holds, both $\text{rank}(\Sigma_W) = K$ and Assumption 1 are necessary for identifying $\Theta^*$.

Proof. The proof is deferred to Appendix B.1.

Since the condition $E[W|X] = 0$ holds under many interesting cases, such as the structured equation model (1.2) and the multivariate Gaussian model for $(Z, X)$, part (2) of Proposition 1 shows that the identifiability of $\Theta^*$ needs to be studied under both Assumption 1 and $\text{rank}(\Sigma_W) = K$. We thus assume $\text{rank}(\Sigma_W) = K$ throughout the paper, that is the covariance matrix $\Sigma_W$ of $W = Z - (A^*)^T X$ is strictly positive definite. In practice, this is also a reasonable assumption as the hidden variable $Z$ usually contains information that cannot be perfectly explained by a linear combination of the observable feature $X$.

To show the sufficiency of Assumption 1 and $\text{rank}(\Sigma_W) = K$, we first describe our procedure of identifying $\Theta^*$, which is constructive, in the following three steps:

(1) identify the coefficient matrix $F^* = \Theta^* + A^* B^*$ in (1.4);

(2) identify $\Sigma_e := \text{Cov}(\varepsilon)$ with $\varepsilon = (B^*)^T W + E$ and use it to construct $P_{B^*}$, the projection matrix onto the row space of $B^*$;

(3) identify $\Theta^*$ from $(I_m - P_{B^*}) Y$. 

8
Recall that $W = Z - (A^*)^T X$ is independent of $E$. In step (2), a key observation from model (1.4) is that, under homoscedastic error, the covariance matrix of $\varepsilon = B^* W + E$ satisfies
\begin{equation}
\Sigma_\varepsilon = (B^*)^T \Sigma_W B^* + \Sigma_E = (B^*)^T \Sigma_W B^* + \tau^2 I_m.
\end{equation}
Since $\text{rank}(B^*) = K$ and $\Sigma_W$ has full rank, (2.1) implies that the row space of $B^*$ coincides with the space spanned by the first $K$ eigenvectors of $\Sigma_\varepsilon$ with non-increasing eigenvalues. We thus propose to identify $P_{B^*}$ via the eigenspace of $\Sigma_\varepsilon$. Under Assumption 1, step (3) uses
\begin{equation}
P_{B^*} Y = \left(\Theta^* P_{B^*} \right)^T X + \left( B^* P_{B^*} \right)^T Z + P_{B^*} E = (\Theta^*)^T X + P_{B^*} E,
\end{equation}
where we write $P_{B^*} = I_m - P_{B^*}$. This further implies
\begin{equation}
\Theta^* = \left[\text{Cov}(X)\right]^{-1} \text{Cov} \left( X, P_{B^*} Y \right).
\end{equation}
The following proposition summarizes the identifiability of $\Theta^*$ under the homoscedastic error.

**Proposition 2.** Under model (1.1) or equivalently (1.4), assume that $\text{rank}(\Sigma_W) = K$, $\Sigma_E = \tau^2 I_m$ and Assumption 1 hold. Then $\Theta^*$ is identifiable from $\text{Cov}(X), \text{Cov}(Y)$ and $\text{Cov}(X,Y)$.

**Proof.** The proof is deferred to Appendix B.2.

Combining Propositions 1 and 2 concludes that Assumption 1 and $\text{rank}(\Sigma_W) = K$ are sufficient and necessary for identifying $\Theta^*$ under homoscedastic error.

### 2.2 Estimation

Recall that we observe the data matrices $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times m}$. Our estimation procedure follows the same steps as in the analysis of model identifiability: (1) first estimate $X F^*$; (2) then estimate $\Sigma_\varepsilon$ and $P_{B^*}$; (3) finally estimate $\Theta^*$.

#### 2.2.1 Estimation of $X F^*$

Recall from model (1.4) that $F^* = \Theta^* + L^*$ where $L^* := A^* B^*$ is a dense matrix and $\Theta^*$ is the row-wise sparse matrix satisfying (1.5). We propose to estimate $F^*$ by $\hat{F} = \hat{\Theta} + \hat{L}$ where $\hat{\Theta}$ and $\hat{L}$ are obtained by solving the following optimization problem
\begin{equation}
(\hat{\Theta}, \hat{L}) = \arg \min_{\Theta, L} \frac{1}{n} \|Y - X (\Theta + L)\|_F^2 + \lambda_1 \|\Theta\|_{\ell_1/\ell_2} + \lambda_2 \|L\|_F^2
\end{equation}
with some tuning parameters $\lambda_1, \lambda_2 \geq 0$. Our estimator is designed to recover both the sparse matrix $\Theta^*$, via the group-lasso regularization (Yuan and Lin, 2006), and the dense matrix $L^*$, via the multivariate ridge regularization. Since our goal in this step is to estimate the best linear predictor $X F^*$, we do not impose the orthogonality constraint (Assumption 1) between $\hat{\Theta}$ and $\hat{L}$ for computational convenience. Computationally, solving (2.3) is efficient with almost the same complexity of solving a group-lasso problem. Specifically, we have the following lemma.
Lemma 1. Let $(\hat{\Theta}, \hat{L})$ be any solution of (2.3), and denote

$$P_{\lambda_2} = X (X^T X + n\lambda_2 I_p)^{-1} X^T, \quad Q_{\lambda_2} = I_n - P_{\lambda_2}$$  \hspace{1cm} (2.4)$$

for any $\lambda_2 \geq 0$ such that $P_{\lambda_2}$ exists. Then $\hat{\Theta}$ is the solution of the following problem

$$\hat{\Theta} = \arg \min_{\Theta} \frac{1}{n} \left\| Q_{\lambda_2}^{1/2} (Y - X\Theta) \right\|^2_F + \lambda_1 \|\Theta\|_{\ell_1/\ell_2},$$  \hspace{1cm} (2.5)$$

and $\hat{L} = (X^T X + n\lambda_2 I_p)^{-1} X^T (Y - X\hat{\Theta})$, where $Q_{\lambda_2}^{1/2}$ is the principal matrix square root of $Q_{\lambda_2}$. Moreover, we have

$$X\hat{F} = X(\hat{\Theta} + \hat{L}) = P_{\lambda_2} Y + Q_{\lambda_2} X\hat{\Theta}.$$  \hspace{1cm} (2.6)$$

Lemma 1 characterizes the role of the regularization parameters $\lambda_2$ and $\lambda_1$. When $\lambda_2 \to 0$, we have $\hat{F} \approx \hat{\Theta} + (X^T X)^+ X^T (Y - X\hat{\Theta}) = (X^T X)^+ X^T Y$, where $(X^T X)^+$ is the pseudo inverse of $X^T X$. Thus, $\hat{F}$ reduces to the generalized least squares estimator. On the other hand, when $\lambda_2 \to \infty$, we can see that $Q_{\lambda_2} \approx I_n$ and $\hat{L} \approx 0$ whence $\hat{F} \approx \hat{\Theta}$ essentially becomes the group-lasso estimator. Later in Remark 2, we will take a closer look at this phenomenon in terms of the convergence rates of $\|X\hat{F} - XF^*\|_F$ under the orthogonal design. The tuning parameter $\lambda_1$ only appears in (2.5) and its magnitude determines the sparsity level of the group-lasso estimator $\hat{\Theta}$. Lemma 1 also implies that the estimator $(\hat{\Theta}, \hat{L})$ is unique if and only if the solution of the group-lasso problem (2.5) is unique. Even if (2.5) has multiple solutions, we can define $(\hat{\Theta}, \hat{L})$ to be any of the solutions and the resulting best linear predictor $X\hat{F} = X(\hat{\Theta} + \hat{L})$ satisfies the desired deviation bounds in Theorem 3.

In the applications when both $m$ and $p$ are large while $K$ is small, $\lambda^*$ can be also viewed as a low-rank matrix with rank $K$. The common approach of estimating a low-rank matrix is to either impose a rank constraint on the matrix known as the reduced-rank approach (Izenman, 2008) or regularize its nuclear norm. The latter is known as the convex relaxation of the reduced-rank approach. We emphasize that, under model (1.4), our approach with the ridge penalty has both theoretical and computational advantages over the aforementioned rank penalized methods. We defer to Section 3.4 for a technical comparison.

Finally, we comment that our method (2.3) can be viewed as the multivariate generalization of the lava approach proposed by Chernozhukov et al. (2017); see also Čevid et al. (2018). Lava estimates the sum of a sparse vector $\beta$ and a dense vector $b$ in linear regression problem $y = X(\beta + b) + \epsilon$ by minimizing the least squares loss plus the penalty $\lambda_1 \|\beta\|_1 + \lambda_2 \|b\|_2^2$. As explained in Chernozhukov et al. (2017), lava is intrinsically different from the elastic net as lava penalizes both $\beta$ and $b$ and the estimate of $(\beta + b)$ is non-sparse, whereas elastic net uses the penalty $\lambda_1 \|\beta\|_1 + \lambda_2 \|b\|_2^2$ and typically yields a sparse estimate of $\beta$. These differences naturally extend to our multivariate setting.

2.2.2 Estimation of $P_B^*$

In this section, we discuss how to estimate the projection matrix $P_B^*$. Consider the singular value decomposition $B^* = V D U^T$, where $V \in \mathbb{R}^{K \times K}$ and $U \in \mathbb{R}^{m \times K}$ are the left and right
singular vectors of $B^*$ and $D \in \mathbb{R}^{K \times K}$ is the diagonal matrix of the non-increasing singular values. It is easily seen that $P_{B^*} = UU^T$. Recall that, from (2.1), $U$ also coincides with the first $K$ eigenvectors of $\Sigma_e$ up to an orthogonal matrix. We thus propose to first estimate $\Sigma_e$ by

$$\hat{\Sigma}_e = \frac{1}{n} (Y - X\hat{F})^T (Y - X\hat{F})$$

(2.7)

with $\hat{F}$ obtained from (2.3) and then estimate $P_{B^*}$ by $\hat{P}_{B^*} = \hat{U}\hat{U}^T$, where $\hat{U}$ consists of the eigenvectors of $\hat{\Sigma}_e$ corresponding to the $K$ largest eigenvalues. We assume $K$ is known for now and defer to Section 5.1 for detailed discussions of selecting $K$.

### 2.2.3 Estimation of $\Theta^*$

After estimating $P_{B^*}$ by $\hat{P}_{B^*}$, motivated by (2.2), we propose to estimate $\Theta^*$ by

$$\tilde{\Theta} = \arg \min_{\Theta} \frac{1}{n} \|Y (I_m - \hat{P}_{B^*}) - X\Theta\|_F^2 + \lambda_3 \|\Theta\|_{\ell_1/\ell_2}$$

(2.8)

with some tuning parameter $\lambda_3 > 0$. Solving the problem in (2.8) is equivalent to solving a group-lasso problem with the projected response matrix $Y(I_m - \hat{P}_{B^*})$.

For the reader’s convenience, we summarize our procedure, Hidden Variable adjustment Estimation (HIVE), in Algorithm 1.

---

**Algorithm 1** The HIVE procedure for estimating $\Theta^*$.

**Require:** Data matrices $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^{n \times m}$, rank $K$, tuning parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$.

1. Estimate $X\hat{F}$ with $\hat{F} = \tilde{\Theta} + \hat{L}$ by solving (2.3).
2. Obtain $\hat{\Sigma}_e$ from (2.7).
3. Compute $\hat{P}_{B^*} = \hat{U}\hat{U}^T$ where $\hat{U}$ are the first $K$ eigenvectors of $\hat{\Sigma}_e$.
4. Estimate $\Theta^*$ by $\tilde{\Theta}$ obtained from (2.8).

---

### 3 Statistical guarantees

In this section, we provide theoretical guarantees for our estimation procedure. In our theoretical analysis, the design matrix $X$ is considered to be deterministic and the analysis can be done similarly for random design by first conditioning on $X$. Recall from model (1.4) that $W$ is only uncorrelated with $X$. To simplify the analysis under the fixed design, we assume the independence between $X$ and $W$ in order to control the deviation of their cross product. We expect that the same theoretical guarantees hold under $\text{Cov}(X,W) = 0$ by using more tedious arguments. We start from the following assumptions on the error matrices $W \in \mathbb{R}^{n \times K}$ and $E \in \mathbb{R}^{n \times m}$.

**Assumption 2.** Let $\gamma_w$ and $\gamma_e$ denote some positive constants.

(1) Assume $\left\{\Sigma_W^{-1/2}W_i\right\}_{i=1}^n$ are i.i.d. $\gamma_w$ sub-Gaussian random vectors\(^1\), where $\Sigma_W = \text{Cov}(W_i)$.

---

\(^1\)A random vector $X \in \mathbb{R}^d$ is $\gamma$ sub-Gaussian if $\langle u, X \rangle$ is $\gamma$ sub-Gaussian for any $\|u\|_2 = 1$.  

---
(2) For any fixed $1 \leq j \leq p$, $\{E_{ij}\}_{i=1}^n$ are i.i.d. $\gamma_e$ sub-Gaussian\(^2\). For any fixed $1 \leq i \leq n$, $\{E_{ij}\}_{j=1}^m$ are independent.

Since part (2) of Assumption 2 does not assume $E_{ij}$ are identically distributed across $1 \leq j \leq m$, this assumption is applicable to both homoscedastic and heteroscedastic errors, provided that $\max_{1 \leq j \leq p} \text{Var}(E_{ij}) \leq \gamma_e^2$. We assume $\Sigma_E = \text{Cov}(E) = \tau^2 I_m$ throughout this section and the heteroscedastic case is discussed in Section 4.

### 3.1 Statistical guarantees of estimating $XF^*$

To establish theoretical properties for $XF^*$ obtained from (2.5), we first generalize the design impact factor of $X$ in Chernozhukov et al. (2017) to the multivariate regression setup. Denote $\tilde{X} = Q_{\lambda_2}^{1/2}X$, where $Q_{\lambda_2}$ is defined in (2.4). For notational simplicity, we suppress the dependence of $\tilde{X}$ on $\lambda_2$. For any constant $c > 0$ and deterministic matrix $\Theta_0 \in \mathbb{R}^{p \times m}$, define the design impact factor as

$$\kappa_1(c, \Theta_0, \lambda_1, \lambda_2) := \inf_{\Delta \in \mathcal{R}(c, \Theta_0, \lambda_1, \lambda_2)} \frac{\|\tilde{X}\Delta\|_F/\sqrt{n}}{\|\Theta_0\|_{\ell_1/\ell_2} - \|\Theta_0 + \Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2}}, \quad (3.1)$$

where

$$\mathcal{R}(c, \Theta_0, \lambda_1, \lambda_2) = \begin{cases} \{\Delta \in \mathbb{R}^{p \times m} \setminus \{0\} : \|\tilde{X}\Delta\|_F/\sqrt{n} \leq 2\lambda_1 \left(\|\Theta_0\|_{\ell_1/\ell_2} - \|\Theta_0 + \Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2}\right)\} \end{cases}. \quad (3.2)$$

It is well known that when $p > n$ the matrix $\tilde{X}^T\tilde{X}$ is singular and the least squares loss is not strictly convex. The design impact factor $\kappa_1(c, \Theta_0, \lambda_1, \lambda_2)$ is introduced to characterize the minimum curvature of the least squares loss in (2.5) when the matrix $\Delta$ is restricted in a feasible set $\mathcal{R}(c, \Theta_0, \lambda_1, \lambda_2)$. It generalizes the widely used Restricted Eigenvalue (RE) condition in high-dimensional regression (Bickel et al., 2009) and is more suitable for prediction (Belloni et al., 2014; Chernozhukov et al., 2017). We refer to Remark 1 for its connection with the RE condition.

Define the following quantity which characterizes the total variation of the multivariate regression in (1.4),

$$V_e = \text{tr}(\Gamma_e), \quad \text{with} \quad \Gamma_e := \gamma_w^2 B^* \Sigma W B^* + \gamma_e^2 I_m, \quad (3.3)$$

where $\text{tr}(\cdot)$ stands for the trace. Let $r_e(\Gamma_e) = \text{tr}(\Gamma_e)/\|\Gamma_e\|_{op}$ denote the effective rank of $\Gamma_e$. Write $M = n^{-1}X^TQ_{\lambda_2}^2X$ and $\hat{\Sigma} = n^{-1}X^TX$. Recall that $P_{\lambda_2}$ and $Q_{\lambda_2}$ are defined in (2.4). The following theorem provides the deviation bounds of $\|X\hat{F} - XF^*\|_F$.

**Theorem 3.** Under model (1.4) and Assumption 2, choose

$$\lambda_1 = 4 \sqrt{\max_{1 \leq j \leq p} M_{jj}} \left(1 + \sqrt{\frac{2\log(p/e')}{r_e(\Gamma_e)}}\right) \sqrt{\frac{V_e}{n}}, \quad (3.4)$$

---

\(^2\)A centered random variable $X$ is $\gamma$ sub-Gaussian if it satisfies $\mathbb{E}[\exp(tX)] \leq \exp(\gamma^2 t^2/2)$ for all $t \geq 0$. 

12
for some $\epsilon' > 0$ and any $\lambda_2 \geq 0$ in (2.3) such that $P_{\lambda_2}$ exists. With probability $1 - \epsilon - \epsilon'$,

$$\frac{1}{n} \left\| X \hat{F} - X F^* \right\|_F^2 \leq \inf_{(\Theta_0, L_0): \Theta_0 + L_0 = F^*} \left[ \frac{2}{n} \left\| X (\hat{L} - L_0) \right\|_F^2 + \frac{2}{n} \left\| Q_{\lambda_2} X (\hat{\Theta} - \Theta_0) \right\|_F^2 \right]$$

$$\leq \inf_{(\Theta_0, L_0): \Theta_0 + L_0 = F^*} \left[ 4 \text{Rem}_1 + 16 \left\| Q_{\lambda_2} \right\|_{op} \cdot \text{Rem}_2 (L_0) + 8 \left\| Q_{\lambda_2} \right\|_{op} \cdot \text{Rem}_3 (\Theta_0) \right],$$

where $\left\| Q_{\lambda_2} \right\|_{op} \leq 1$ and

$$\text{Rem}_1 = \left( \sqrt{\text{tr}(P_{\lambda_2}^2)} + \sqrt{2 \log(m/\epsilon) \left\| P_{\lambda_2}^2 \right\|_{op}} \right)^2 \frac{V_2}{n},$$

$$\text{Rem}_2 (L_0) = \lambda_2 \cdot \text{tr} \left[ L_0^T \Sigma (\hat{\Sigma} + \lambda_2 I_p)^{-1} L_0 \right],$$

$$\text{Rem}_3 (\Theta_0) = \lambda_2^2 \cdot [\kappa_1 (1/2, \Theta_0, \lambda_1, \lambda_2)]^{-2}.$$

**Proof.** The proof is deferred to Appendix B.3. \qed

Since Assumption 1 is not assumed in Theorem 3, neither $L^*$ nor $\Theta^*$ can be identified individually. Nevertheless, our estimator $\hat{F}$ minimizes the error over all possible combinations of $\Theta_0$ and $L_0$ satisfying $\Theta_0 + L_0 = F^*$. In particular, it holds for $(\Theta^*, L^*)$ whenever they are identifiable. As expected from (2.3), the prediction error comes from two sources: estimating $L^*$ from the multivariate ridge regression and estimating $\Theta^*$ from the group-lasso. Specifically, $\text{Rem}_1$ and $\text{Rem}_2 (L_0)$ are, respectively, the variance and bias terms from the ridge regression while $\text{Rem}_3 (\Theta_0)$ corresponds to the estimation error of the group-lasso. In the following, we comment on these three terms one by one.

**Remark 1** (Design impact factor and $\lambda_1$). The remainder term $\text{Rem}_3 (\Theta_0)$ depends on the design impact factor $\kappa_1 (c, \Theta_0, \lambda_1, \lambda_2)$ and the tuning parameter $\lambda_1$. We first discuss the connection of $\kappa_1 (c, \Theta_0, \lambda_1, \lambda_2)$ and the Restricted Eigenvalue (RE) condition of $X$ defined as

$$\kappa(s, \alpha) = \min_{S \subset \{1, \ldots, p\}, |S| \leq s} \min_{\Delta \in \mathbb{C}(s, \alpha)} \frac{\left\| X \Delta \right\|_F}{\sqrt{n} \| \Delta \|_F},$$

where $\alpha \geq 1$ is a constant, $s$ is a positive integer and $\mathbb{C}(s, \alpha) := \{ \Delta \in \mathbb{R}^{p \times m} \setminus \{0\} : \alpha \| \Delta \|_{\ell_1/\ell_2} \geq \| \Delta \|_{\ell_1/\ell_2} \}$. Similarly, denote by $\tilde{\kappa}(s, \alpha)$ the RE condition of $\tilde{X} = Q_{\lambda_2}^{1/2} X$. In Lemma 8 of Appendix C, we show that, for any constant $c \in (0, 1)$,

$$[\kappa_1 (c, \Theta_0, \lambda_1, \lambda_2)]^2 \geq \frac{[\kappa(s_0, \alpha_c)]^2}{(1 + c)^2 s_0} \geq \frac{\lambda_2}{\sigma_1 + \lambda_2} \cdot \frac{[\kappa(s_0, \alpha_c)]^2}{(1 + c)^2 s_0},$$

where $\alpha_c = (1 + c)/(1 - c)$, $s_0 = \| \Theta_0 \|_{\ell_0/\ell_2}$ and $\sigma_1$ is the leading eigenvalue of $\hat{\Sigma} = n^{-1} X^T X$.

The first inequality is proved in Chernozhukov et al. (2017) for $m = 1$. Here we extend it to $m \geq 2$, and we further establish the second inequality which characterizes the relation between $\kappa_1 (c, \Theta_0, \lambda_1, \lambda_2)$ and $\kappa(s_0, \alpha_c)$. It is well known that the RE $\kappa(s_0, \alpha_c)$ is lower bounded by a constant with high probability when the rows of $X$ are i.i.d. sub-Gaussian vectors with $\lambda_{\min} (\text{Cov}(X)) > \epsilon'$ for some constant $\epsilon' > 0$ and $s_0 = O(n)$ (Rudelson and Zhou, 2013).
Together with the second inequality in (3.6), we obtain \( |\kappa_1(c, \Theta_0, \lambda_1, \lambda_2)|^2 \gtrsim \lambda_2/[s_0(\sigma_1 + \lambda_2)] \). Thus, when \( \lambda_2 \) is relatively large comparing to \( \sigma_1, \kappa_1(c, \Theta_0, \lambda_1, \lambda_2) \) scales as \( 1/\sqrt{s_0} \).

Note that the magnitude of \( \text{Rem}_3(\Theta_0) \) also depends on the tuning parameter \( \lambda_1 \), which is further related to the choice of \( \lambda_2 \) via the diagonal entries of \( M = n^{-1}X^TQ_{\lambda_2}^2X \). To further simplify \( \text{Rem}_3(\Theta_0) \), in Lemma 8 of Appendix C we prove

\[
\max_{1 \leq j \leq p} M_{jj} \leq \max_{1 \leq j \leq p} \widehat{\Sigma}_{jj} \left( \frac{\lambda_2}{\sigma_q + \lambda_2} \right)^2,
\]

(3.7)

where \( \sigma_q \) is the smallest non-zero eigenvalue of \( \widehat{\Sigma} \). Combining (3.6) and (3.7), we obtain that

\[
\text{Rem}_3(\Theta_0) \lesssim \frac{\lambda_2(\sigma_1 + \lambda_2)}{\sigma_q + \lambda_2} \max_{1 \leq j \leq p} \left[ \frac{s_0}{|\kappa(s_0, 3)|^2} \right] \left[ 1 + \frac{\log(p/e')}{r_e(\Gamma_{e})} \right] \frac{V_{e}}{n}. \]

The first two remainder terms \( \text{Rem}_1 \) and \( \text{Rem}_2(L_0) \) depend on the choice of \( \lambda_2 \) in a more complicated way. To make the remainder terms more transparent, we can bound them from above via the eigenvalues of \( \widehat{\Sigma} \). To save space, we collect all the results and only present the simplified deviation bounds of \( \|X\widehat{\Theta} - XF^*\|_F^2 \) in the following corollary. Recall that \( \sigma_1 \gtrsim \cdots \gtrsim \sigma_q \) denote the non-zero eigenvalues of \( \widehat{\Sigma} \) with \( q = \text{rank}(X) \).

**Corollary 4.** Suppose conditions of Theorem 3 and Assumption 1 hold. Assume \( \kappa(s_*, 3) > 0 \) where \( \kappa(s_*, 3) \) is defined in (3.5) with \( s_* = \|\Theta^*\|_{0, t_2} \). With probability \( 1 - \epsilon - \epsilon' \),

\[
\frac{1}{n} \left\| X\widehat{\Theta} - XF^* \right\|_F^2 \lesssim \left[ \sum_{k=1}^{q} \left( \frac{\sigma_k}{\sigma_k + \lambda_2} \right)^2 + \left( \frac{\sigma_1}{\sigma_1 + \lambda_2} \right)^2 \log(m/\epsilon) \right] \frac{V_{e}}{n} + \frac{\sigma_1\lambda_2}{(\sigma_1 + \lambda_2)(\sigma_q + \lambda_2)} \lambda_2 \|L^*\|_F^2 + \frac{\sigma_1 + \lambda_2}{\sigma_q + \lambda_2} \lambda_2 \max_{1 \leq j \leq p} \widehat{\Sigma}_{jj} \left[ 1 + \frac{\log(p/e')}{r_e(\Gamma_{e})} \right] \frac{s_*}{|\kappa(s_*, 3)|^2} \frac{V_{e}}{n}.
\]

*Proof.* The proof is deferred to Appendix C.2. \( \square \)

Since Assumption 1 guarantees the identifiability of \( \Theta^* \), we replace with \( (\Theta^*, L^*) \) the infimum over \( (\Theta_0, L_0) \) satisfying \( \Theta_0 + L_0 = F^* \) in Theorem 3.

**Remark 2** (Orthonormal design). To draw connections with the existing results on group-lasso and ridge estimators, we consider the orthonormal design \( \widehat{\Sigma} = I_p \). The deviation bounds in Theorem 7 and Corollary 4 reduce to (after ignoring the logarithmic factors)

\[
\frac{1}{n} \left\| X\widehat{\Theta} - XF^* \right\|_F^2 \lesssim \left( \frac{1}{1 + \lambda_2} \right)^2 \frac{pV_{e}}{n} + \left( \frac{\lambda_2}{1 + \lambda_2} \right)^2 \|L^*\|_F^2 + \left( \frac{\lambda_2}{1 + \lambda_2} \right)^2 \frac{s_*V_{e}}{n}.
\]

(3.8)

The first two terms are the variance and bias due to the ridge penalty while the third term is the error of the group-lasso. As \( \lambda_2 \) increases, the variance term of the ridge gets smaller whereas the bias term of the ridge and the error of group-lasso become larger. Optimizing the right hand side of (3.8) over \( \lambda_2 \) yields

\[
\lambda_2 = \frac{pV_{e}/n}{\|L^*\|_F^2 + s_*V_{e}/n}.
\]

(3.9)
(a) When $L^* = 0$, model (1.4) reduces to $Y = (\Theta^*)^T X + \varepsilon$ with error $\varepsilon = B^T W + E$. (3.9) becomes $\lambda_2 = p/s_* \text{ and max}_j M_{jj} \simeq p^2/(p + s_*)^2$ from (3.7). Consequently, (3.4) implies

$$
\lambda_1 \simeq \left( \frac{p}{s_* + p} \right)^2 \left( 1 + \sqrt{\frac{\log(p/\varepsilon)}{r_{\varepsilon}(\varepsilon)}} \right) \sqrt{\frac{V_\varepsilon}{n}}
$$

and (3.8) reduces to

$$
\frac{1}{n} \|X\hat{\Lambda} - XF^*\|^2_F \lesssim \left( \frac{s_*}{p + s_*} \right)^2 \frac{pV_\varepsilon}{n} + \left( \frac{p}{p + s_*} \right)^2 \frac{s_*V_\varepsilon}{n} \lesssim \frac{s_*V_\varepsilon}{n},
$$

which is the optimal rate of the group-lasso estimator.

(b) When $\Theta^* = 0$, the model (1.4) reduces to $Y = (L^*)^T X + \varepsilon$, and $\lambda_2 = pV_\varepsilon/(n\|L^*\|^2_F)$ from (3.9) and $s_* = 0$. After some simple calculation, (3.8) yields

$$
\frac{1}{n} \|X\hat{\Lambda} - XF^*\|^2_F \lesssim \min \left\{ \frac{pV_\varepsilon}{n}, \|L^*\|_F^2 \right\} \lesssim \sqrt{\frac{pV_\varepsilon}{n}} \|L^*\|_F,
$$

which is the optimal rate of the ridge regression (Hsu et al., 2014).

Combining scenarios (a) and (b), we conclude that the convergence rate (3.8) of our estimator $\hat{\Theta}$ with the optimal tuning parameters $\lambda_1$ and $\lambda_2$ matches the best possible rate even if $L^* = 0$ or $\Theta^* = 0$ were known a priori. For this reason, we refer to our estimator $\hat{\Theta}$ as an adaptive estimator.

One might notice that when $\Theta^* = 0$ and $K$ is much smaller than both $p$ and $m$, we have a multivariate regression with the coefficient matrix $L^*$ exhibiting a low-rank structure. A natural option is to use a reduced-rank estimator to estimate $L^*$. In Section 3.4, we show the advantage of our ridge-type estimator over the reduced-rank estimator under our setting.

### 3.2 Statistical guarantees of estimating $\Theta^*$

Recall that $\Theta^*$ is estimated from (2.8) by using the estimates of the projection matrix $P_{B^*}$. The estimation error of $\bar{\Theta}$ should depend on how accurately one can estimate $P_{B^*}$. We state a general theorem below which establishes the non-asymptotic upper bounds of $\|\bar{\Theta} - \Theta^*\|_1$ for $\bar{\Theta}$ obtained from (2.8) by using any estimator $\hat{P}$ of $P_{B^*}$ in lieu of $\hat{P}_{B^*}$. Let $\Lambda_1$ denote the largest eigenvalue of $B^*^T \Sigma W B^*$.

**Theorem 5.** Under model (1.4) and Assumptions 1 – 2, assume $\kappa(s_*,4) > 0$. Let $\bar{\Theta}$ be any solution of problem (2.8) by using any estimator $\hat{P} \in \mathbb{R}^{m \times m}$ of $P_{B^*}$ in lieu of $\hat{P}_{B^*}$. Choose any $\lambda_3 \geq \bar{\lambda}_3$ in (2.8) with

$$
\bar{\lambda}_3 = 4\gamma_\varepsilon \sqrt{\max_{1 \leq j \leq p} \sqrt{m} + \frac{2\log(p/\varepsilon)}{\sqrt{n}}}. \tag{3.11}
$$

On the event $\{\|\hat{P} - P_{B^*}\|_F \lesssim \xi_n\}$ for some proper sequence $\xi_n$, with probability $1 - \epsilon - 2e^{-cK}$ for some constant $c > 0$, one has

$$
\|\bar{\Theta} - \Theta^*\|_1 \lesssim \max \left\{ \lambda_3, \frac{(\bar{\lambda}_3)^2}{\kappa^2(s_*,4)} \right\} \frac{s_*}{\kappa^2(s_*,4)}. \tag{3.12}
$$
where
\[
\tilde{\lambda}_3 = \left\{ \frac{1}{\sqrt{n}} \| X F^* \|_{op} + \sqrt{\Lambda_1} \left( 1 + \frac{\sqrt{K}}{n} \right) \right\} \frac{\kappa(s_*, 4)}{\sqrt{s_*}} \xi_n. \tag{3.13}
\]

**Proof.** The proof is deferred to Appendix B.4. □

If \( K \) is small, one can replace \( \sqrt{K/n} \) in (3.13) by \( \sqrt{K \log(n)/n} \) and the resulting probability of (3.12) will become \( 1 - \epsilon - 2n^{-cK} \) which, by choosing \( \epsilon = (p \vee n)^{-1} \), tends to one as \( n \to \infty \). The same argument is applicable to the rest of the theorems.

Theorem 5 holds for any estimator \( \hat{P} \) of \( \hat{P}_B^* \) with convergence rate \( \| \hat{P} - \hat{P}_B^* \|_F \lesssim \xi_n \). The effect of \( \hat{P} \) on the estimation error of \( \hat{\Theta} \) is characterized by the term \( (\tilde{\lambda}_3)^2 s_* / [\lambda_3 \kappa^2(s_*, 4)] \) in (3.12) via the choice of \( \lambda_3 \). When \( \hat{P}_B^* \) can be estimated very accurately, for instance when \( B^* \) is known, \( \xi_n \) is fast enough such that \( \tilde{\lambda}_3 \leq \lambda_3 \). We can take \( \lambda_3 = \hat{\lambda}_3 \) to obtain the convergence rate \( \tilde{\lambda}_3 s_* / \kappa^2(s_*, 4) \). We refer to this as the oracle rate since it is the optimal rate for estimating \( \Theta^* \) from \( Y P_B^\dagger = X \Theta^* + E P_B^\dagger \) when \( B^* \) is known (cf. Lounici et al. (2011)). On the other hand, when \( \hat{P} \) has a slow rate such that \( \hat{\lambda}_3 > \lambda_3 \), one needs to take a larger \( \lambda_3 \) to achieve the best trade-off between the two terms in (3.12). It is easy to see that in this scenario the optimal \( \lambda_3 \) is equal to \( \hat{\lambda}_3 \) and the resulting convergence rate is \( \tilde{\lambda}_3 s_* / \kappa^2(s_*, 4) \).

**Remark 3** (On the benefit of group-lasso). As seen above, when \( \tilde{\lambda}_3 \leq \lambda_3 \), the convergence rate (3.12) reduces to the oracle rate \( \lambda_3 s_* / \kappa^2(s_*, 4) \). Moreover, if \( \log(p) = o(m) \) and \( \max_j \hat{\Sigma}_{jj} = O(1) \) hold, (3.11) implies that \( \tilde{\lambda}_3 = O(\sqrt{m/n}) \) by choosing \( \epsilon = p^{-1} \). As a result, provided that \( [\kappa(s_*, 4)]^{-1} = O(1) \), the average error per response satisfies \( \sum_{j=1}^{p} \left[ m^{-1} \sum_{e=1}^{m} (\hat{\Theta}_{je} - \Theta^*_{je})^2 \right]^{1/2} = m^{-1/2} \| \hat{\Theta} - \Theta^* \|_{\ell_1/\ell_2} = O(s_*/\sqrt{n}) \), which does not depend on logarithmic factors of the feature dimension \( p \) and is faster than the standard rate of the lasso applied to each column of \( Y \) separately. Such a phenomenon is known as the benefit of group-lasso (Lounici et al., 2011). Recall that we also use the group-lasso in our first step (2.3) for estimating \( X F^* \). This benefit of group-lasso remains and can be seen from the choice of \( \lambda_1 \) in (3.4). Indeed, if \( \log(p) = o(r_e(\Gamma_\epsilon)) \) holds, by choosing \( \epsilon = p^{-1} \) in (3.4), the \( \log(p) \) term in \( \lambda_1 \) is negligible. The quantity \( r_e(\Gamma_\epsilon) \) is the effective rank of \( \Gamma_\epsilon \) and it depends on the interplay of \( B^* T \Sigma_W B^* \) and \( \Sigma_E \). If \( \| B^* T \Sigma_W B^* \|_{op} \) is small (e.g., upper bounded by a constant), then \( r_e(\Gamma_\epsilon) \asymp m \), whereas if \( \lambda_1(B^* T \Sigma_W B^*) \asymp \lambda_K(B^* T \Sigma_W B^*) \asymp m \), we have \( r_e(\Gamma_\epsilon) \asymp K \).

In the following theorem, we establish non-asymptotic upper bounds of the estimation error of our estimator \( \hat{P}_B^* \) obtained from Section 2.2.2. The proof is based on a variant of the Davis-Kahan theorem (Yu et al., 2014) together with careful control of the estimation error of \( \hat{\Sigma}_e \). Let \( \Lambda_K \) denote the \( K \)th largest eigenvalue of \( (B^*)^T \Sigma_W B^* \).

**Theorem 6.** Under model (1.4) and Assumptions 1 – 2, assume \( \kappa(s_*, 4) > 0 \) and \( m \leq e_n \). For some constants \( c, c' > 0 \), one has
\[
P \left\{ \| \hat{P}_B^* - P_B^* \|_F \leq c \cdot \text{Rem}(P_B^*) \right\} \geq 1 - \epsilon' - 5m^{-c'},
\]
where, with $V_ε$ and $Γ_e$ defined in (3.3),

$$Rem(P_{B*}) = \frac{1}{Λ_K} \left\{ V_ε \sqrt{\frac{\log m}{n}} + \frac{λ_2 σ_1}{λ_2 + σ_1} ||L^*||_F^2 + \frac{σ_k V_ε}{n} \right\} + \frac{λ_2(σ_1 + λ_2)}{(σ_q + λ_2)^2} \max_{1 ≤ j ≤ p} \hat{Σ}_{jj} \left( 1 + \frac{log(p/ε)}{r_ε(Γ_e)} \right) \frac{s_* V_ε}{n} \right\}. $$

(3.14)

**Proof.** The proof is deferred to Appendix B.5. □

Recall that $\hat{P}_{B*}$ relies on the estimates of $Σ_e$ and $XF^*$. The term $V_ε \sqrt{\log m/n}$ is the oracle error of estimating $Σ_e$ in Frobenius norm even if $XF^*$ were known. The other three terms in $Rem(P_{B*})$ come from the errors of estimating $XF^*$ in Corollary 4.

When $\hat{Θ}$ is obtained from (2.8) by using $\hat{P}_{B*}$, combining Theorem 5 and Theorem 6 yields the final rate of $||\hat{Θ} - Θ^*||_{Γ_1/Γ_2}$ with explicit dependency on all quantities. To simplify its expression, we now assume some conditions. Without loss of generality, we first assume that the design matrix is standardized such that $\hat{Σ}_{jj} = 1$ for all $1 ≤ j ≤ p$.

**Assumption 3.**

(a) $||\hat{Σ}_{S,S}||_{op} = O(1)$, $|κ(s_*, 4)|^{-1} = O(1)$.

(b) $Λ_1 ∼ Λ_K ∼ m$, where $Λ_1$ and $Λ_K$ are the first and $K$th eigenvalues of $(B^*)^T Σ_W B^*$;

(c) $\frac{1}{n} ||XL^*||^2_{op} = O(m)$, $||Θ^*||^2_{op} = O(m + s_*)$.

The verification of Assumption 3 is deferred to Section 3.3. Under Assumption 3, the following Corollary 7 simplifies the rates of $||\hat{Θ} - Θ^*||_{Γ_1/Γ_2}$ in Theorem 5. For two sequences $a_n$ and $b_n$, we write $a_n ≍ b_n$ for $a_n = O(b_n)$ up to a logarithmic factor.

**Corollary 7.** Assume conditions of Theorem 5 and Assumption 3 hold. Further assume $K = O(n)$. With probability tending to one, there exists a suitable choice of $λ_2$ in (2.3) such that

$$\frac{1}{\sqrt{m}} ||\hat{Θ} - Θ^*||_{Γ_1/Γ_2} ≍ \max \left\{ \frac{s_*}{\sqrt{n}}, \sqrt{\frac{s_* (m + s_*)}{m}} \cdot Err(P_{B*}) \right\} $$

(3.15)

where

$$Err(P_{B*}) = \min \left\{ σ_1 \frac{||L^n||^2}{m} + \frac{Ks_*}{n}, \frac{qK}{\sqrt{n}}, \sqrt{\frac{||L^n||^2_{op}}{m} \cdot (p + σ_1 s_*)/n} + \frac{Ks_*}{n} \right\} + \frac{K}{\sqrt{n}}.$$

**Proof.** The proof is deferred to Appendix C.3. □

In view of (3.15), $s_*/\sqrt{n}$ is the oracle rate for estimating $Θ^*$ as discussed after Theorem 5. The term $Err(P_{B*})$ quantifies the minimum price to pay for estimating $P_{B*}$ over all choices of $λ_2$ and it is the minimum of three error terms which are related with the estimation of $XF^*$. In order to facilitate understanding, we further simplify (3.15) in two particular settings.

**Remark 4** (Further simplified rates in low- and high-dimensional settings).

17
(i) Suppose $p = \text{rank}(X) < n$, $p \asymp s_*$, $\sigma_1 = O(1)$ and $K = O(\sqrt{p} \wedge \sqrt{n/p} \wedge m)$. Then (3.15) becomes

\[
\frac{1}{\sqrt{m}} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{p}{\sqrt{n}}, \quad \text{when } p = O(m);
\]

\[
\frac{1}{\sqrt{m}} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{p}{\sqrt{n}} + \left(\frac{p}{\sqrt{n}}\right)^2, \quad \text{when } m = O(1).
\]

Note that the oracle rate in this case is $p/\sqrt{n}$. As long as $p/\sqrt{n} = o(1)$ which is the minimum requirement for consistent estimation of $\Theta^*$ in $\ell_1/\ell_2$ norm, our estimator $\tilde{\Theta}$ achieves the oracle rate. If $m$ grows at least of order $p$, we also allow $K$ to grow but no faster than $\sqrt{p} \wedge \sqrt{n/p} \wedge m$.

(ii) Suppose $p \geq n$, $s_* < n$ and $K = O(\sqrt{s_*} \wedge \sqrt{n/s_*} \wedge m)$. Then (3.15) becomes

\[
\frac{1}{\sqrt{m}} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_*}{\sqrt{n}} + \sqrt{s_*} \frac{\|L^*\|_F}{\sqrt{m}} \cdot \min \left\{ \sigma_1 \frac{\|L^*\|_F}{\sqrt{m}}, \sqrt{\frac{(p + \sigma_1 s_*)K}{n}} \right\}, \quad \text{when } s_* = O(m);
\]

\[
\frac{s_*}{\sqrt{n}} + \left(\frac{s_*}{\sqrt{n}}\right)^2 + s_* \|L^*\|_F \cdot \min \left\{ \sigma_1 \frac{\|L^*\|_F}{\sqrt{m}}, \sqrt{\frac{(p + \sigma_1 s_*)}{n}} \right\}, \quad \text{when } m = O(1).
\]

In high-dimensional case, the dimension $m$ plays a more significant role. When $m$ is fixed, one needs $s_* \sigma_1 \|L^*\|_F^2 = o(1)$ or $s_* \|L^*\|_F(p + \sigma_1 s_*)^{1/2} = o(n^{1/2})$ for estimation consistency. This requirement is much more relaxed as $m$ tends to infinity. The benefit of a large $m$ can be viewed as the blessing of dimensionality. In the sequel, we focus on $s_* = O(m)$, which leads to the following two sub-cases:

\[
\frac{1}{\sqrt{m}} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_*}{\sqrt{n}} + 2 \frac{s_\sigma_1 \|L^*\|_F^2}{m}, \quad \text{if } \frac{\|L^*\|_F^2}{m} \leq \frac{(p + \sigma_1 s_*)K}{n \alpha^2};
\]

\[
\frac{1}{\sqrt{m}} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_*}{\sqrt{n}} + \sqrt{s_\sigma_1 (p + \sigma_1 s_*)K \|L^*\|_F^2} m \cdot \frac{\|L^*\|_F^2}{m}, \quad \text{if } \frac{\|L^*\|_F^2}{m} \geq \frac{(p + \sigma_1 s_*)K}{n \alpha^2}.
\]

Intuitively, the first case is more likely to occur if $\sigma_1$, the largest eigenvalue of $\hat{\Sigma}$, has moderate magnitude, such as $\sigma_1 = O(p/n)$. We refer to Section 3.3 for more comments on this order of $\sigma_1$. In this case, assuming $m \asymp n^\alpha$ for some constant $\alpha \geq 1/2$, the rate matches the oracle rate $s_* \sqrt{n}$ if $\sigma_1 \|L^*\|_F^2 = O(\sqrt{s_\alpha n^{2\alpha-1}})$. The larger $\alpha$ is, the weaker the requirement on $\|L^*\|_F^2$ becomes. On the other hand, when $\hat{\Sigma}$ has spiked eigenvalues, for instance $\sigma_1 \asymp p$, the second case is more likely to occur. In this case, assuming $\sigma_1 \asymp p$ and $K = O(1)$, our estimator $\tilde{\Theta}$ achieves the oracle rate $s_* \sqrt{n}$ if $\|L^*\|_F^2 = O(m/p)$. In Section 3.3, we provide examples to justify the condition $\|L^*\|_F^2 = O(m/p)$.

When $\Theta^*$ is identifiable, $\hat{\Theta}$ obtained in (2.3) can be viewed as an initial estimator of $\Theta^*$. In fact, the convergence rate of $\|\hat{\Theta} - \Theta^*\|_{\ell_1/\ell_2}$ is established in Lemma 7 of Appendix B.6. In the following remark, we elaborate the improvement of our final estimator $\hat{\Theta}$ over this initial estimator $\tilde{\Theta}$ in terms of their convergence rates. Empirical comparisons of these two estimators are considered in our simulation of Section 6.
Remark 5 (Comparison of \(\hat{\Theta}\) and \(\tilde{\Theta}\)). Assume conditions of Corollary 4 and Assumption 3 hold. With suitable choices of \(\lambda_1\) and \(\lambda_2\), \(\hat{\Theta}\) defined in (2.3) satisfies

\[
\frac{1}{\sqrt{m}} \|\hat{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_\ast \sqrt{K}}{\sqrt{n}} + \sqrt{s_\ast} \sqrt{\frac{s_\ast \|L^*\|_F^2}{m}}.
\]

Comparing this rate to (3.15), the advantage of \(\hat{\Theta}\) over the initial estimator \(\tilde{\Theta}\) is substantial. For instance, in the low-dimensional case (i) of Remark 4, we have \(m^{-1/2} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{p}{\sqrt{n}}\) provided that \(p/\sqrt{n} = o(1)\). In contrast, \(m^{-1/2} \|\hat{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{p(K/n)^{1/2}}{n}\) which has an extra \(K^{1/2}\) factor, even if \(\sigma_1 \|L^*\|_F^2 / m\) is sufficiently small. In the high-dimensional case (ii), one has

\[
\frac{1}{\sqrt{m}} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_\ast \sqrt{m}}{\sqrt{n}} + \frac{s_\ast (m + s_\ast)}{m} \left( \frac{\sigma_1 \|L^*\|_F^2}{m} + \sqrt{\frac{s_\ast}{n}} \right)
\]

which is always faster than (3.16) provided that \(\sigma_1 \|L^*\|_F^2 = o(m/s_\ast)\) and \(s_\ast = O(mK)\). Note that in (3.16) we need \(\sigma_1 \|L^*\|_F^2 = o(m/s_\ast)\) for the consistency of \(\tilde{\Theta}\). For further illustration, suppose \(s_\ast = O(m), m \asymp n^\alpha\) for some \(\alpha \geq 1/2\) and \(\sigma_1 \|L^*\|_F^2 = O(\sqrt{s_\ast n^{2\alpha - 1}})\), then \(m^{-1/2} \|\tilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_\ast \sqrt{m}}{\sqrt{n}}\) corresponds to the oracle rate, whereas (3.16) becomes \(m^{-1/2} \|\hat{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \frac{s_\ast (K/n)^{1/2}}{n} + s_\ast n^{(\alpha - 1)/2}\) which may even diverge to infinity.

### 3.3 Validity of Assumption 3 and conditions in Remark 4

In this section we provide theoretical justifications for Assumption 3 as well as some conditions in Remark 4.

Part (a) of Assumption 3 contains standard conditions on the design matrix. The validity of \([\kappa(s_\ast, 4)^{-1} = O(1)\) is already discussed in Remark 1. To show \(\|\hat{\Sigma}_{S,S_\ast}\|_{op} = O(1)\), assume the rows of \(X \Sigma^{-1/2}\) are i.i.d. sub-Gaussian random vectors with bounded sub-Gaussian constant, where \(\Sigma = \text{Cov}(X)\). Then provided that \(s_\ast = O(n)\), one has \(\|\hat{\Sigma}_{S,S_\ast} - \Sigma_{S,S_\ast}\|_{op} = O_p(\|\Sigma_{S,S_\ast}\|_{op} \sqrt{s/n})\) (see for instance, Vershynin (2012)), which implies \(\|\hat{\Sigma}_{S,S_\ast}\|_{op} = O_p(1)\) when \(\|\Sigma_{S,S_\ast}\|_{op} = O(1)\). Moreover, when \(\|\Sigma\|_{op} = O(1)\), the condition \(\sigma_1 = O_{p}(p/n)\) in part (ii) of Remark 4 is guaranteed since \(\|\hat{\Sigma} - \Sigma\|_{op} = O_p(\sqrt{p/n} \vee (p/n))\) by Vershynin (2012). To show \(\sigma_1 \asymp p\) in part (ii) of Remark 4, suppose \(\|\Sigma\|_{op} = O(p)\) and \(\log(p) = o(n)\). Since \(\|\hat{\Sigma} - \Sigma\|_{op} \leq \|\hat{\Sigma} - \Sigma\|_{F} = O_p(\sqrt{p \log(p)/n})\) (for instance, see the argument in Lemma 16 of Appendix C.5), one can deduce that \(\sigma_1 \asymp p\).

Condition (b) is standard when \(m\) (and also \(K\)) is fixed. When \(m\) grows with \(n\), we note that \(\varepsilon = WB^* + E\) follows a factor model where \(W\) is the matrix of \(K\) stochastic factors and \(B^*\) is the factor loading matrix. Condition (b) is known as the pervasiveness assumption in the factor model literature for identification and consistent estimation of the row space of the factor loading \(B^*\) (Chamberlain and Rothschild, 1983; Connor and Korajczyk, 1986; Bai and Ng, 2008; Bai, 2003; Fan et al., 2011, 2013, 2017). In particular, condition (b) holds if \(c \leq \lambda_K(\Sigma_W) \leq \lambda_1(\Sigma_W) \leq C\) for some constants \(c, C > 0\), and the columns of \(B^*\) are i.i.d. copies of a \(K\)-dimensional sub-Gaussian random vector whose covariance matrix has bounded eigenvalues. It is worth mentioning that this assumption is only used to simplify the order of
In (3.13) and Rem($P_{B^*}$) in (3.14). If $\Lambda_1$ and $\Lambda_K$ have different rates, we can replace them by corresponding rates and simplify the error bound of $\tilde{\Theta}$ analogously.

For condition (c), since $\|\Theta^*\|_{l_0/l_2} \leq s_*$, $\|\Theta^*\|_{op} = O(m + s_*)$ holds when either $\|\Theta^*\|_{\infty} = O(1)$ or entries of $\Theta^*$ are i.i.d. samples from a mean-zero distribution with bounded fourth moment (Bai and Yin, 1993). Condition $\frac{1}{n}\|XL^*\|_{op}^2 = O(m)$ requires the dense signal (see below for more interpretation) is not too large. In Lemma 10 of the Appendix, when $\Sigma^{-1/2}X_i$ are i.i.d. sub-Gaussian random vectors with bounded sub-Gaussian constant and under part (b) of Assumption 3, we show that $\frac{1}{n}\|XL^*\|_{op}^2 = O_p(m)$ holds provided that $K = O(n), \lambda_1(\Sigma_Z) = O(1)$ and $[\lambda_K(\Sigma_W)]^{-1} = O(1)$.

In the end, we comment on the magnitude of $\|L^*\|_F^2$ as it appears in the rate of $\|\tilde{\Theta} - \Theta^*\|_{l_1/l_2}$. Since $L^* = A^*B^*$ with $A^* = \{E[XX^T]\}^{-1}E[XZ^T]$, we can interpret $L^*$ as the effect of the hidden variables $Z$ on the response that can be explained by a linear combination of the observed $X$. In the mediation analysis via structural equation models, $L^*$ is known as the indirect effect. Under model (1.4), the estimation of the non-sparse coefficient matrix $\Theta^* + L^*$ becomes more challenging in high dimension when $\|L^*\|_F^2$ is large, and the estimation error is further accumulated in the final estimator $\tilde{\Theta}$ in Corollary 7. Thus, intuitively $\|L^*\|_F^2$ cannot grow too fast in order to guarantee the consistency of $\tilde{\Theta}$. This can be compared to the standard results in linear regression. For instance, in linear regression $y = X\beta + \epsilon$ where $y \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ is dense, one needs $\|\beta\|_2^2 = o(1)$ for consistent estimation when $p > n$; see Hsu et al. (2014); Dicker (2016) for the minimax lower bound.

Finally, it is of interest to derive under what conditions $\|L^*\|_F^2$ is small and how small it can be. By part (b) of Assumption 3, we first have $\|L^*\|_F^2/m \leq \|A^*\|_F^2\|B^*\|_{op}/m = O(\|A^*\|_F^2)$ provided that $c \leq \lambda_K(\Sigma_W) \leq \lambda_1(\Sigma_W) \leq C$ for some constants $c, C > 0$. Since $A^* = \Sigma^{-1/2}\text{Cov}(X, Z)$ by assuming, without loss of generality, $X$ and $Z$ have zero means, intuitively $\|A^*\|_F$ is small when either (1) $\text{Cov}(X, Z)$ is close to zero or (2) $\Sigma$ is large.

To show when case (1) holds, suppose the smallest eigenvalue of $\Sigma$ is lower bounded by a positive constant. When $\text{Cov}(X, Z)$ is very sparse with $\|\text{Cov}(X, Z)\|_{l_0} = O(1)$ and $\max_{j,k} |\text{Cov}(X_j, Z_k)| \leq \xi$, one has $\|A^*\|_2^2 = O(\xi^2)$ which could vanish if $\xi = o(1)$. When $\text{Cov}(X, Z)$ is dense with $\|\text{Cov}(X, Z)\|_{l_0} \geq c'(pK)$ for some small constant $c' > 0$, if the range of the nonzero entries of $\text{Cov}(X, Z)$ is bounded, an application of Pólya-Szegő inequality (see, for instance, Dragomir (2015)) yields $\|\text{Cov}(X, Z)\|_F \leq \|\text{Cov}(X, Z)\|_{l_1/l_1} = O(1)$, one has $\|L^*\|_F^2/m = O(\|A^*\|_2^2) = O(1/(pK))$.

To show when case (2) holds, we consider the setting that $X$ follows an approximate factor model $X = GF + W'$, where the noise $W'$ and the factor $F$ are independent, $\text{Cov}(F)$ and $\text{Cov}(W')$ have bounded eigenvalues and the loading matrix $\Gamma \in \mathbb{R}^{p \times K}$ satisfies the perservasiveness assumption $\lambda_K(\Gamma \Gamma^T) \gtrsim p$ for some $1 \leq K \leq p$. We refer to the third paragraph of this section for further discussion of the perservasiveness assumption. In this scenario, $\Sigma = \Gamma \text{Cov}(F) \Gamma^T + \text{Cov}(W')$ has $K$ spiked eigenvalues with order at least $p$. To show the order of $\|L^*\|_F^2$, we consider the eigen-decomposition of $\Sigma = \sum_{j=1}^p d_j v_j v_j^T$ with $d_1 \geq \cdots \geq d_p$. Further write $V_{(K)} = \sum_{j=1}^K d_j$ in (3.13) and Rem($P_{B^*}$) in (3.14). If $\Lambda_1$ and $\Lambda_K$ have different rates, we can replace them by corresponding rates and simplify the error bound of $\tilde{\Theta}$ analogously.
\[(v_1, \ldots, v_K) \in \mathbb{R}^{p \times K} \text{ and } V_{(-K)} = (v_{K+1}, \ldots, v_p) \in \mathbb{R}^{p \times (p-K)}. \] Provided that
\[\left\| V_{(K)}^T \text{Cov}(X, Z) \right\|_{op} \leq c/\sqrt{p} \quad \text{and} \quad \left\| V_{(-K)}^T \text{Cov}(X, Z) \right\|_{op} \leq c/\sqrt{p}, \tag{3.17}\]
for some sufficiently small constant \(c > 0\), we obtain
\[\|A^*\|_{op} = \|\Sigma^{-1}\text{Cov}(X, Z)\|_{op} \leq \frac{1}{d_K} \left\| V_{(K)}^T \text{Cov}(X, Z) \right\|_{op} + \frac{1}{d_p} \left\| V_{(-K)}^T \text{Cov}(X, Z) \right\|_{op} = O(1/\sqrt{p}).\]
Moreover, if \(c \leq \lambda_K(\Sigma_Z) \leq \lambda_1(\Sigma_Z) \leq C\) for some constants \(c, C > 0\), by noting that \(\Sigma_W = \Sigma_Z - [\text{Cov}(X, Z)]^T\Sigma^{-1}\text{Cov}(X, Z)\), one can deduce \(c/2 \leq \lambda_K(\Sigma_W) \leq \lambda_1(\Sigma_W) \leq C\), whence, \(\|L^*\|^2_{21}/m = O(\|A^*\|^2_{21}) = O(K/p)\). The condition (3.17) requires that: (1) the order of \(\|\text{Cov}(X, Z)\|_{op}\) can not be greater than \(\sqrt{p}\); (2) the columns of \(\text{Cov}(X, Z)\) and \(V_{(-K)}\) are approximately orthogonal. From a practical perspective, under the structural equation model (1.2), condition (3.17) implies that the causal effect of \(X\) on \(Z\) (i.e., \(A^*\)) is weak due to the spiked eigenvalues of \(\Sigma\). However, by introducing the factor model for \(X\), the association between the hidden variable \(Z\) and the factor \(F\) is enhanced and thus the indirect effect of \(X\) on \(Y\) via \(Z\) cannot be ignored.

### 3.4 Comparison with the reduced-rank estimator

In this section, we state our reasoning for using the ridge penalty in (2.3) rather than the commonly used low-rank approach. In particular, we compare our estimator (2.3), or equivalently the multivariate ridge regression, with the reduced-rank estimator under our model (1.4), \(Y = (L^*)^TX + (B^*)^TW + E\) when \(\Theta^* = 0\). Since \(L^* = A^*B^*\) exhibits a low-rank structure when both \(p\) and \(m\) are relatively large comparing to \(K\), one could estimate \(L^*\) by the reduced-rank estimator (Izenman, 1975, 2008; Bunea et al., 2011)
\[\hat{L}^{(RR)} = \arg\min_{L} \|Y - XL\|_F^2 + \mu \cdot \text{rank}(L) \tag{3.18}\]
for some tuning parameter \(\mu > 0\).

There have been extensive research on the estimation of a low-rank matrix in both regression and matrix completion settings, for instance, Bing and Wegkamp (2019); Bunea et al. (2011, 2012); Candès and Plan (2011); Candès and Tao (2009); Rohde and Tsybakov (2011); Reinsel and Velu (1998); Giraud (2011, 2015); Obozinski et al. (2011); Negahban and Wainwright (2011); Koltchinskii et al. (2011); Yuan et al. (2007), a list that is far from exhaustive. In regression setting, the reduced-rank estimator \(\hat{L}^{(RR)}\) is shown to be optimal for both prediction and estimation in high-dimensional settings (Bunea et al., 2011). Despite of its popularity, we give several reasons why the multivariate ridge-type estimator is more suitable in our setting.

First, we focus on the comparison of the convergence rate of prediction error. Suppose part (b) of Assumption 3 holds. From Corollary 4, by using \(s_e = 0\), \(V_e = O(Km)\) and \(\sum_k \sigma_k \leq q_0 \sigma_1\) with \(q = \text{rank}(X)\), one can deduce that our estimator \(\hat{X}F\) satisfies (notice that \(F^* = L^*\) as \(\Theta^* = 0\))
\[\frac{1}{n} \left\| X\hat{F} - XL^* \right\|_F^2 \lesssim \frac{\sigma_1 qKm}{\lambda_2 n} + \frac{\sigma_1 Km \log(m/\epsilon)}{\lambda_2 n} + \lambda_2 \|L^*\|_{21}^2.\]
When \( \log(m) = O(q) \), choosing \( \epsilon = m^{-1} \) and optimizing the above rate over \( \lambda_2 \) yield
\[
\frac{1}{n} \left\| X \hat{F} - XL^* \right\|_F^2 \lesssim \sqrt{\frac{\sigma_1 K q m}{n}} \|L^*\|_F.
\]
On the other hand, Bunea et al. (2011) showed that the reduced-rank estimator \( \hat{L}^{(RR)} \) in (3.18) has the following prediction error
\[
\frac{1}{n} \left\| X \hat{L}^{(RR)} - XL^* \right\|_F^2 \lesssim \frac{K}{n} \|WB^* + E\|_{op}^2 \lesssim \frac{K m}{n} \left\| (WB^* + E) \Sigma^{-1/2}_\epsilon \right\|_{op}^2 \lesssim \frac{K(n + m)m}{n}
\]
where the first inequality can be proved in the same way as Koltchinskii et al. (2011); Bunea et al. (2011), the second inequality holds because the operator norm of \( \Sigma_\epsilon = (B^*)^T \Sigma_W B^* + \tau^2 I_m \) is of order \( m \) under part (b) of Assumption 3, and the last inequality is due to the deviation bounds of the operator norm of random matrices whose rows are i.i.d. sub-Gaussian random vectors (Vershynin, 2012). Clearly, by \( \|L^*\|_F^2 \leq K \|L^*\|_{op}^2 \) and recalling that \( q = \text{rank}(X) \leq \min\{n, p\} \), the rate of our estimator \( X \hat{F} \) is potentially faster if
\[
\sigma_1 \|L^*\|_{op}^2 \leq \frac{nm}{q} \left(1 + \frac{m}{n}\right)^2
\]
which usually holds, especially when \( m \) is larger than \( n \) (see the discussion on \( \|L^*\|_F \) in Section 3.3).

To the best of our knowledge, the reduced-rank estimator only achieves the minimax rate when the error covariance matrix has a bounded operator norm (Koltchinskii et al., 2011; Rohde and Tsybakov, 2011). However, as shown above, the error in our model (1.4) has covariance matrix \( (B^*)^T \Sigma_W B^* + \tau^2 I_m \) whose operator norm is order of \( m \). In this case, the reduced-rank estimator is no longer optimal. Instead, our ridge type estimator leverages the fact that \( \|L^*\|_F \) is small and leads to a much faster prediction rate. This theoretical comparison is further corroborated by simulation studies in Appendix D.

Finally, from computational perspective, if we replace \( \lambda_2 \|L\|_F^2 \) in (2.3) by the rank penalty \( \lambda_2 \cdot \text{rank}(L) \), the resulting optimization becomes non-convex and is computationally challenging. For computational convenience, one may instead penalize the nuclear norm of \( L \) rather than its rank such that the resulting optimization problem can be solved by the ADMM algorithm. However, to establish statistical guarantees for the corresponding estimator, on top of the RE condition for the group-lasso penalty, this approach may require additional conditions on the design matrix due to the nuclear norm regularization (Koltchinskii et al., 2011). In contrast, the proposed estimator (2.3) can be computed efficiently and has much weaker restrictions on the design matrix for provable guarantees.

4 Extension to heteroscedastic noise

We have discussed the identifiability and estimation in model (1.1) when the errors are homogeneous. In practice, the multivariate response \( Y \) may correspond to measurement of different properties (e.g., phenotypes) whose values could differ in scales. To deal with this problem, in this section we extend the model by allowing heteroscedastic errors, \( \Sigma_E = \text{diag}(\tau_1^2, \ldots, \tau_m^2) \), and discuss how to modify our approach to account for this heteroscedasticity.
4.1 Identifiability

Recalling the identifiability in Section 2.1, we observe that the heteroscedasticity only affects the identification of $P_{B^\ast}$ in step (2). When $\Sigma_E = \text{diag}(\tau_1^2, \ldots, \tau_m^2)$ one has

$$\Sigma_e = (B^\ast)^T \Sigma_W B^\ast + \text{diag}(\tau_1^2, \ldots, \tau_m^2).$$

(4.1)

In contrast to the homoscedastic case, the eigenspace of $\Sigma_e$ corresponding to the first $K$ eigenvalues, in general, no longer coincides with the row space of $B^\ast$. Consequently, we cannot identify $P_{B^\ast}$ via the eigenspace of $\Sigma_e$ as in Section 2.1. To overcome this difficulty, we resort to a newly developed procedure called HeteroPCA proposed by Zhang et al. (2018). For completeness, we restate their procedure in Algorithm 2. The main idea is to iteratively perform the singular value decomposition (SVD) on the off-diagonal elements of $\Sigma_e$ to impute its diagonal. Under a mild incoherence condition on the row space of $B^\ast$, $P_{B^\ast}$ can be recovered by applying Algorithm 2 to $\Sigma_e$ and as a result, $\Theta^\ast$ is identifiable from (2.2). We summarize the identifiability in Proposition 8 below.

Recall that $P_{B^\ast} = UU^T$ with $U := U(K) \in \mathbb{R}^{m \times K}$ being the first $K$ right singular vectors of $B^\ast$, and $\Lambda_1$ and $\Lambda_K$ are the first and $K$th eigenvalues of $(B^\ast)^T \Sigma_W B^\ast$, respectively. Let $\{e_1, \ldots, e_m\}$ denote the canonical basis of $\mathbb{R}^m$.

**Proposition 8.** Under model (1.1) or equivalently (1.4) and Assumption 1, assume $\Sigma_E = \text{diag}(\tau_1^2, \ldots, \tau_m^2)$ and $\text{rank}(\Sigma_W) = K$. Further assume

$$\frac{\Lambda_1}{\Lambda_K} \max_{1 \leq j \leq m} \|e_j^T U\|_2^2 \leq C_U$$

(4.2)

for some constant $C_U > 0$. Then $P_{B^\ast}$ can be uniquely determined via Algorithm 2 with input $\hat{\Sigma} = \Sigma_e$, $r = K$ and some sufficiently large number of iterations $T$. As a result, $\Theta^\ast$ is identifiable.

An application of Theorem 3 in Zhang et al. (2018) guarantees the recovery of $P_{B^\ast}$ from $\Sigma_e$ and the rest of the proof follows the same lines as the proof of Proposition 2. Compared to the homoscedastic case, we need an extra condition (4.2) for identifying $P_{B^\ast}$, which can be viewed as the price to pay for allowing heteroscedasticity. Inherent from the HeteroPCA algorithm, this condition is to rule out matrices $U$ that are well aligned with canonical basis vectors. Otherwise, one cannot separate $(B^\ast)^T \Sigma_W B^\ast$ from a diagonal matrix. We also note that $\frac{\max_{1 \leq j \leq m} \|e_j^T U\|_2^2}{\text{rank}}$ is known as the incoherence constant in the matrix completion literature (Candès and Tao, 2009; Candès et al., 2011). When $\Lambda_1 \asymp \Lambda_K$, $\max_{1 \leq j \leq m} \|e_j^T U\|_2^2 = O(1)$ in (4.2) is much weaker than the typical incoherence condition $\max_{1 \leq j \leq m} \|e_j^T U\|_2^2 = O(K/m)$, assumed in the matrix completion literature. Finally, Proposition 3 in Zhang et al. (2018) implies that condition (4.2) in general cannot be further relaxed in order to recover $P_{B^\ast}$ from $\Sigma_e$.

**Remark 6** (Identification via PCA when $m \to \infty$). We propose to use HeteroPCA to identify $P_{B^\ast}$ in the presence of heteroscedasticity since it guarantees the identifiability of $\Theta^\ast$ for any $m \geq K$ under the condition (4.2). Directly applying PCA to $\Sigma_e$ as in Section 2.1 may not recover
$P_{B^*}$ hence not identify $\Theta^*$. However, we remark that PCA is robust against the departure from homoscedasticity and even the diagonal structure of $\Sigma_E$ when $\Lambda_K$, the $K$th eigenvalue of $(B^*)^T \Sigma W B^*$, diverges fast enough as $m \to \infty$. Specifically, at the population level, applying PCA to $\Sigma$ will recover $P_{B^*}$ asymptotically provided that
\[
\sqrt{K}\|\Sigma_E\|_{op} = o(\Lambda_K), \quad \text{as } m \to \infty.
\]
A sufficient condition would be $\Lambda_K \gtrsim m$ and $\sqrt{K}\|\Sigma_E\|_{op} = o(m)$. This phenomenon is known as the blessing of dimensionality in the factor model literature (Bai, 2003; Fan et al., 2013, 2017). Most of the methods in surrogate variable analysis, for instance Lee et al. (2017); McKennan and Nicolae (2019), rely on this robustness of PCA. Their methods thus only guarantee the asymptotic identification of $\Theta^*$ when $m \to \infty$, and are not applicable if $m$ is fixed.

**Algorithm 2** HeteroPCA($\hat{\Sigma}, r, T$)

1: Input: matrix $\hat{\Sigma}$, rank $r$, number of iterations $T$.
2: Set $N_{ij}^{(0)} = \hat{\Sigma}_{ij}$ for all $i \neq j$ and $N_{ii}^{(0)} = 0$.
3: for $t = 1, \ldots, T$ do
4: Calculate SVD: $N(t) = \sum_i \lambda_i^{(t)} u_i^{(t)} (v_i^{(t)})^T$, where $\lambda_1^{(t)} \geq \lambda_2^{(t)} \geq \cdots \geq 0$.
5: Let $\tilde{N}(t) = \sum_{i=1}^r \lambda_i^{(t)} u_i^{(t)} (v_i^{(t)})^T$.
6: Set $N_{ij}^{(t+1)} = \hat{\Sigma}_{ij}$ for all $i \neq j$ and $N_{ii}^{(t+1)} = \tilde{N}_i^{(t)}$.
7: Output $U(T) = [u_1^{(T)}, \ldots, u_r^{(T)}]$.

**4.2 Estimation**

Our estimation procedure remains the same except estimating $U$ by HeteroPCA in Algorithm 2. To be specific, we consider the estimator $\hat{P}_{B^*} = \hat{U} \hat{U}^T$, where $\hat{U}$ is obtained from Algorithm 2 with the input $\hat{\Sigma} = \hat{\Sigma}_e$, $r = K$ and a large $T$ for the algorithm to converge. Our simulation reveals that $T = 5$ usually yields satisfactory results. We still assume $K$ is known and defer the discussion of selecting $K$ to Section 5.1. We state the modified algorithm in Algorithm 3, named as Heteroscedastic Hidden Variable adjustment Estimation (H-HIVE).

**Algorithm 3** The H-HIVE procedure for estimating $\Theta^*$.

**Require:** Data matrices $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^{n \times m}$, rank $K$, number of iterations $T$, tuning parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$.

1: Estimate $X \hat{F}$ with $\hat{F} = \hat{\Theta} + \hat{L}$ by solving (2.3).
2: Obtain $\hat{\Sigma}_e$ from (2.7).
3: Compute $\hat{P}_{B^*} = \hat{U} \hat{U}^T$ where $\hat{U}$ is obtained from HeteroPCA($\hat{\Sigma}_e, K, T$) in Algorithm 2.
4: Estimate $\Theta^*$ by solving (2.8) with $\hat{P}_{B^*}$ in lieu of $\hat{P}_{B^*}$.
4.3 Statistical guarantees

Our estimation algorithm enjoys similar statistical guarantees as in Section 3. First, since $\hat{F}$ is the same estimator as obtained from (2.3), the deviation bounds of $\|X\hat{F} - XF^*\|_F$ in Theorem 3 and Corollary 4 still hold under Assumption 2. Second, Theorem 9 below provides non-asymptotic upper bounds for $\|\tilde{P}_{B^*} - P_{B^*}\|_F$ where $\tilde{P}_{B^*} = \tilde{U}\tilde{U}^T$ with $\tilde{U}$ obtained from Algorithm 2. Finally, since $\tilde{\Theta}$ is obtained from the same criterion in (2.8) by using $\tilde{P}_{B^*}$ in place of $\tilde{P}_{B^*}$, the convergence rate of $\|\hat{\Theta} - \Theta^*\|_{\ell_1/\ell_2}$ immediately follows by the theorem below in conjunction with Theorem 5.

**Theorem 9.** Under the same conditions of Theorem 6, assume condition (4.2) holds and $\text{Rem}(P_{B^*}) \leq c\sqrt{K}$ for some constant $c > 0$ with $\text{Rem}(P_{B^*})$ defined in (3.14). For some constants $\epsilon', \epsilon'' > 0$, the estimator $\tilde{P}_{B^*} = \tilde{U}\tilde{U}^T$ with $\tilde{U}$ obtained from Algorithm 2 satisfies

$$\mathbb{P}\left\{\|\tilde{P}_{B^*} - P_{B^*}\|_F \leq \epsilon' \cdot \text{Rem}(P_{B^*})\right\} \geq 1 - \epsilon' - 5m^{-\epsilon''}.$$  

**Proof.** The proof is deferred to Appendix B.7. \qed

The proof of Theorem 9 mainly relies on a new robust $\sin \Theta$ theorem stated in Appendix A, which provides upper bounds for the Frobenius norm of $\sin \Theta(\tilde{U}, U) := \tilde{U}^TU$, where $\tilde{U}$ is the output of Algorithm 2 and $U_\perp$ is its orthogonal complement. The new $\sin \Theta$ theorem complements Theorem 3 in Zhang et al. (2018) which controls the operator norm of $\sin \Theta(\tilde{U}, U)$. In order to establish the rate of $\hat{\Theta}$, we need this new result to control the Frobenius norm of the estimated eigenspace. This technical tool can be of its own interest and potentially useful for many other problems.

The validity of Theorem 9 also hinges on the condition $\text{Rem}(P_{B^*}) \leq c\sqrt{K}$. Under conditions of Corollary 7 and Remark 4, by inspecting their proofs in Appendix C.3, one can verify that $\text{Rem}(P_{B^*}) = O(\sqrt{K})$ holds for a suitable choice of $\lambda_2$ provided that, up to a logarithmic factor, $p\sqrt{K} = O(n)$ in the low-dimensional case or $(s_\ast \vee \sqrt{K})\sqrt{K} = O(n)$ and $\sigma_1\|L^*\|_F^2 = O(m\sqrt{K})$ in the high-dimensional case.

**Remark 7** (Effect of heteroscedasticity on the rate of $\|\hat{\Theta} - \Theta\|_{\ell_1/\ell_2}$). Heteroscedasticity affects the estimation error of $\hat{\Theta}$ implicitly via $V_\varepsilon$ defined in (3.3) and $\tilde{\lambda}_3$ in (3.11). For simplicity of presentation, we assumed $E_{ij}$ shares the same sub-Gaussian constant $\gamma_\varepsilon$ for $1 \leq j \leq m$ in Assumption 2. To illustrate the effect of heteroscedasticity, one could instead assume $E_{ij}/\tau_j$ is $\gamma_\varepsilon$ sub-Gaussian for $1 \leq j \leq m$, then by inspecting the proof and using modified arguments in Lemmas 11 – 14 in Appendix C.5, it is straightforward to show that the same results in Theorems 3, 5, 6 and 9 hold with $V_\varepsilon$ and $\tilde{\lambda}_3$ replaced by

$$V_\varepsilon' = \gamma_\varepsilon^2 \text{tr}\left(B^*T\Sigma\Lambda B^*\right) + \gamma_\varepsilon^2 m_\varepsilon^2, \quad \tilde{\lambda}_3 = 4\gamma_\varepsilon\tilde{\tau} \sqrt{\max_{1 \leq \ell \leq p} \sum_{j=1}^m \tau_j^2 \left(\tau_j^2 + \sqrt{2\log(p/\epsilon)}\right)/\sqrt{n}}.$$  

where $\tilde{\tau}^2 = m^{-1}\sum_{j=1}^m \tau_j^2$. The quantity $\tilde{\tau}^2$ reduces to $\tau^2$ in the homoscedastic case. But in the presence of strong heteroscedasticity, $\tilde{\tau}^2$ can be of order different from $O(1)$.
To conclude this section, we compare the estimation error of $\tilde{P}_{B^*}$ in Theorem 9 with the estimator $\hat{P}_{B^*}$ by using PCA obtained in Section 2.2.2 under the heteroscedastic case.

**Theorem 10.** Suppose the same conditions of Theorem 6 hold. Then for some constants $c, c' > 0$, one has

$$\Pr\left\{\|\hat{P}_{B^*} - P_{B^*}\|_F \leq c \cdot \text{Rem}^{(h)}(P_{B^*})\right\} \geq 1 - c' - 5m^{-c'}$$

where

$$\text{Rem}^{(h)}(P_{B^*}) = \text{Rem}(P_{B^*}) + \frac{1}{\Lambda_K} \sum_{j=1}^{m} (\tau_j^2 - \bar{\tau}_j^2)^2 \right) \right)^{1/2}$$

with $\text{Rem}(P_{B^*})$ defined in (3.14) and $\Lambda_K$ being the $K$th eigenvalue of $B^{*T}\Sigma W B^*$.

**Proof.** The proof is deferred to Appendix B.8. \qed

Comparing (4.3) with (3.14), the last term in (4.3) is the bias of PCA induced by the heteroscedasticity as it zero when the error $E$ is homoscedastic. In general, the bias term vanishes if $\Lambda_K$ is large and the degree of heteroscedasticity is small, such as $\Lambda_K \gtrsim m$ and $\sum_{j=1}^{m} (\tau_j^2 - \bar{\tau}_j^2)^2 = O(m)$ as $m \to \infty$. This can be viewed as the sample analog of the robustness of PCA that we mentioned in Remark 6. However, we note that, even if the bias term converges to 0, it may have a slower rate than $\text{Rem}(P_{B^*})$ which renders the estimation error of $\hat{P}_{B^*}$ larger than that of $\tilde{P}_{B^*}$.

5 Practical considerations

In this section, we address several practical concerns. First, we consider how to select $K$, the number of hidden variables. Then, we discuss the effect of overestimating/underestimating $K$ on the estimation of $\Theta^*$. Selection of tuning parameters and recommendation of standardization are discussed subsequently.

5.1 Selection of $K$

Recall that $\varepsilon = WB^* + E$ and $K$ corresponds to the rank of the unknown coefficient matrix $B^*$. When $\varepsilon$ and $W$ are both observable, estimation of the rank of coefficient matrix has been studied by Bunea et al. (2011, 2012); Giraud (2011); Bing and Wegkamp (2019) in the framework of multivariate regression. However, since $\varepsilon$ and $W$ are both unobserved, we view $\varepsilon = WB^* + E$ as a factor model with $K$ being the number of factors. Bai and Ng (2002) proposed information based criterion to select $K$. However, both this approach and the aforementioned ones in the regression setting require to know the noise level quantified by $\|\Sigma E\|_{op}$. While it might be possible to estimate $\Sigma E$ in view of (4.1), the theoretical justification of this class of methods is complicated under our model.
In this paper, we consider an eigenvalue ratio approach originally developed by Lam and Yao (2012); Ahn and Horenstein (2013) in factor models. Specifically, we estimate \( \varepsilon \) by \( \hat{\varepsilon} = Y - X\hat{F} \) with \( \hat{F} \) obtained from (2.3) and construct \( \hat{\Sigma}_e \) as (2.7). We then propose to estimate \( K \) by

\[
\hat{K} = \arg \max_{j \in \{1, 2, \ldots, K\}} \frac{\hat{\lambda}_j}{\hat{\lambda}_{j+1}},
\]

where \( \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \) are the eigenvalues of \( \hat{\Sigma}_e \) and \( \hat{K} \) is a pre-specified number, for example, \( \hat{K} = [(n \wedge m)/2] \) (Lam and Yao, 2012). This procedure does not require the knowledge of any unknown quantity, such as the noise level \( \|\Sigma_E\|_{\text{op}} \).

The following theorem provides theoretical justification for the above procedure. The proof is deferred to Appendix B.9. Recall that \( \Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_K \) denote the first \( K \) eigenvalues of \((B^*)^T\Sigma W B^*\), and we allow the noise to be heteroscedastic \( \Sigma_E = \text{diag}(\tau_1^2, \ldots, \tau_m^2) \).

**Theorem 11.** Under model (1.1) or equivalently (1.4), suppose Assumption 2 and condition (b) in Assumption 3 hold. Assume \( \max_{1 \leq j \leq m} \tau_j^2 = O(1) \), \( \text{Rem}(P_{B^*}) = o(1) \) with \( \text{Rem}(P_{B^*}) \) defined in (3.14). Then with probability \( 1 - \epsilon' - 5m^{-c''} \) for some constant \( c'' > 0 \),

\[
\frac{\hat{\lambda}_j}{\hat{\lambda}_{j+1}} = O(1), \quad \text{for } 1 \leq j \leq K - 1, \quad \text{and} \quad \frac{\hat{\lambda}_{K+1}}{\hat{\lambda}_K} = O \left( \text{Rem}(P_{B^*}) + m^{-1} \right).
\]

Under Assumption 3, \( K = O(1) \), \( s_* = o(n) \) and \( \sigma_1 \|L^*\|_F^2 = o(m) \), one can deduce from the proof of Corollary 7 that \( \text{Rem}(P_{B^*}) = o(1) \) for a suitable choice of \( \lambda_2 \). In addition, if \( m \to \infty \), we obtain \( \hat{\lambda}_K / \hat{\lambda}_{K+1} \to \infty \). Thus, the maximizer of \( \hat{\lambda}_j / \hat{\lambda}_{j+1} \) is no smaller than \( K \) asymptotically, i.e., \( \hat{K} \geq K \), which partially justifies the criterion in (5.1).

The criterion (5.1) is also related to the “elbow” approach, which is often used to determine the number of principle components for PCA. If we plot the ratio \( \hat{\lambda}_j / \hat{\lambda}_{j+1} \) against \( j \), by Theorem 11 and the above discussion we expect to see that the curve has a sharp increase at \( j = K \), giving an angle in the graph. We can then select this value \( j \) as an estimate of \( K \). In our simulation, this simple elbow approach and the criterion (5.1) usually give us the same results. In Section 6.2, we conduct simulations to compare our criterion (5.1) with some other existing methods for selecting \( K \) such as Buja and Eyuboglu (1992).

### 5.2 Consequence of overestimating or underestimating \( K \)

It is of interest to understand the effect of selecting an incorrect \( K \) on the estimation of \( \Theta^* \). Recall that, after estimating \( K \) by \( \hat{K} \), we construct \( \hat{P}_K = \hat{U}_K \hat{U}_K^T \) and use it in lieu of \( \hat{P}_{B^*} \) in (2.8) to estimate \( \Theta^* \). For illustration purpose, we consider the case that \( \hat{K} = r \) for some positive integer \( 1 \leq r \leq m \). At the population level, assume we know the orthogonal matrix \( U_r = (u_1, \ldots, u_r) \) such that when \( r < K \), \( U_r \) is simply the first \( r \) columns of \( U := U_K \), the right singular vectors of \( B^* \), and when \( r \geq K \), the first \( K \) columns of \( U_r \) align with those of \( U \) and the rest of \( r - K \) columns are arbitrary but orthogonal to \( U \). Similar to \( P_{B^*} = U_K U_K^T \), \( P_r = U_r U_r^T \) is also a projection matrix. Write \( P_r^\perp = I_m - U_r U_r^T \). The following lemma demonstrates that the effect of using \( P_r^\perp \) to estimate \( \Theta^* \) is characterized by the difference of \( P_r \) and \( P_{B^*} \).
Lemma 2. Under model (1.4) and Assumption 1, one has

\[ P_r^T Y = \begin{cases} 
|\Theta^* + A^* B^* (P_{B^*} - P_r)|^T X + P_r^T (B^*)^T W + P_r^T E, & \text{if } r < K; \\
(\Theta^*)^T X - (P_r - P_{B^*})(\Theta^*)^T X + P_r^T E, & \text{if } r > K; \\
(\Theta^*)^T X + P_r^T E, & \text{if } r = K.
\]

As we can see, if \( r < K \), the estimand of (2.8) is \( \Theta^* + A^* B^* (P_{B^*} - P_r) = \Theta^* + A^* (B^*)_{(-r)} \) where we apply SVD to \( B^* = \sum_j d_j u_j v_j^T \) with \( d_j \) being non-increasing singular values and \( (B^*)_{(-r)} = \sum_{j>r} d_j u_j v_j^T \). Thus, the estimator in (2.8) has bias \( A^* (B^*)_{(-r)} \). Intuitively, if the last \( K-r \) singular values of \( B^* \), \( d_{r+1},...,d_K \), are relatively small and close to zero, we expect the bias \( A^* (B^*)_{(-r)} \) to be negligible. In this case, underestimating \( K \) may still lead to a reasonably accurate estimate of \( \Theta^* \). On the other hand, if \( r > K \), our estimator is also biased, and the bias equals to \( -(P_r - P_{B^*})(\Theta^*)^T = -P_r (\Theta^*)^T \) (the equality holds by Assumption 1). Its magnitude depends on the angle between \( \Theta^* \) and the last \( r-K \) columns of \( U_r \).

5.3 Choosing tuning parameters \( \lambda_1, \lambda_2 \) and \( \lambda_3 \)

Recall that our procedure (Algorithms 1 and 3) require three tuning parameters \( \lambda_1, \lambda_2 \) and \( \lambda_3 \). Since the first two parameters \( (\lambda_1, \lambda_2) \) and the third one \( \lambda_3 \) appear in two optimization problems (2.3) and (2.8), respectively, we propose to select \( (\lambda_1, \lambda_2) \) and \( \lambda_3 \) separately by cross validation. When estimating \( F^* \) in (2.3), we can search \( \lambda_1 \) and \( \lambda_2 \) over a two-way grid to minimize the mean squared prediction error via \( k \)-fold cross validation. Similarly, when estimating \( \Theta^* \) in (2.8), we can tune \( \lambda_3 \) by \( k \)-fold cross validation over a grid of \( \lambda_3 \). We set \( k = 10 \) in our simulation.

5.4 Standardization

In steps (2.3) and (2.8) of our estimation procedure, the tuning parameters \( \lambda_1 \) and \( \lambda_3 \) depend on \( \max_{1 \leq j \leq p} \hat{\Sigma}_{jj} \) from Theorems 3 and 5. This dependency comes from the union bounds argument for controlling \( \max_{1 \leq j \leq p} \|X_j^T P_{\lambda_2} \|_2 \). To tighten the bound in practice, we recommend standardizing the columns of \( X \) to unit variance. Since the means of \( Y \) and \( X \) do not affect the estimation of \( \Theta^* \), one can also center both \( X \) and \( Y \) before fitting the model.

6 Simulation study

Data generating mechanism. We set \( K = s_3 = 3 \) throughout the simulation settings. The design matrix is sampled from \( X_i \sim N_p(0, \Sigma) \) for \( 1 \leq i \leq n \) where \( \Sigma_{j\ell} = (-1)^{j+\ell} \rho^{j-\ell} \) for all \( 1 \leq j, \ell \leq p \). Under \( Z = A^T X + W \), to generate \( A \) and \( B \), we sample \( A_{jk} \sim \eta \cdot N(0.5, 0.1) \) and \( B_{k\ell} \sim N(0.1, 1) \) independently for all \( 1 \leq j \leq p \), \( 1 \leq k \leq K \) and \( 1 \leq \ell \leq m \). We use \( \eta \) to control the magnitude of \( A \) hence the dense matrix \( L = AB \). We generate the first \( s_3 \) rows of \( \Theta_{\text{raw}} \) by sampling each entry independently from \( N(\mu_\Theta, \sigma^2_\Theta) \) and set the rest rows to 0. The final \( \Theta \) is chosen as \( \Theta_{\text{raw}} (I_m - B(B^T B)^{-1} B) \) which has the same row sparsity as \( \Theta_{\text{raw}} \) and satisfies Assumption 1. For the error terms, we independently generate \( W_{ik} \sim N(0,1) \) for all \( 1 \leq i \leq n \) and \( 1 \leq k \leq K \). For homoscedastic case, \( E_{ij} \) for \( 1 \leq i \leq n \) and \( 1 \leq j \leq m \) are i.i.d. realizations.
of $N(0, 1)$. For heteroscedastic case, we independently generate $E_{ij} \sim N(0, \tau^2_j)$ where, to vary the degree of heterogeneity, we follow the simulation setting in Zhang et al. (2018) and choose

$$\tau^2_j = \frac{m \cdot v_j^0}{\sum_j v_j^0}, \quad v_1, \ldots, v_m \overset{i.i.d.}{\sim} \text{Unif}[0, 1].$$

This choice of $\tau^2_j$ guarantees $\sum_{j=1}^m \tau^2_j/m = 1$ and $\alpha$ controls the degree of heterogeneity: a larger $\alpha$ corresponds to more heterogeneity.

Methods. We consider both HIVE and H-HIVE in Algorithms 1 and 3. All tuning parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$ are chosen via 10-fold cross validation as described in Section 5.3. We set the number of iterations $T = 5$ for H-HIVE, as the estimator converges quickly in our simulation.

Depending on the setting, we compare our method with estimators from the following list:

- Oracle: the estimator from (2.8) by using $P_B = B^T(B^T B)^{-1}B$ with the true $B$.
- Lasso: the group-lasso estimator from R-package glmnet.
- Ridge: the multivariate ridge estimator from R-package glmnet.
- HIVE-init: $\hat{\Theta}$ obtained from solving (2.3) in step (1) of Algorithm 1.
- SVA: the surrogate variable analysis summarized in the following three steps: (i) compute $\hat{\Theta}_{LS} = (X^T X)^{-1}X^T Y$; (ii) obtain $\hat{P}$ by the first $K$ right singular vectors of $Y - X \hat{\Theta}_{LS}$; (iii) estimate $\Theta^*$ by $\hat{\Theta}_{LS}(I_m - \hat{P})$.
- OLS: the ordinary least squares estimator $\hat{\Theta}_{LS} = (X^T X)^{-1}X^T Y$.

The Oracle estimator requires the knowledge of true $B$ and is used as a benchmark to show the effect of $\hat{P}$ on the estimation of $\Theta^*$ in (2.8). We also consider HIVE-init, which is used as an initial estimator in Algorithms 1 and 3, to illustrate the improvement of HIVE (H-HIVE) via (2.8). There are many variants of SVA in the literature, for instance, Wang et al. (2017); Lee et al. (2017); McKennan and Nicolae (2019). Since they have similar performances in our setting, we only consider the aforementioned one.

To make fair comparison, we provide the true $K$ for SVA, HIVE and H-HIVE in Section 6.1. We then show the performance of selecting $K$ by using the criterion (5.1) and the permutation test by Buja and Eyuboglu (1992) in Section 6.2.

6.1 Comparison with existing methods

In this section, we compare the performance of Oracle, Lasso, Ridge, SVA, HIVE-init, HIVE and H-HIVE in three different settings: (1) small $p$ and small $m$ ($m = p = 20$); (2) small $p$ and large $m$ ($m = 150, p = 20$); (3) large $p$ and small $m$ ($m = 20, p = 150$). For each setting, we fix $n = 100$ and consider both homoscedastic and heteroscedastic cases.

We choose $\mu_{\Theta} = 3$ and $\sigma_{\Theta} = 0.1$ and vary $\rho \in \{0, 0.5\}$ across all settings. For the homoscedastic case we vary $\eta \in \{0.1, 0.3, 0.5, \ldots, 1.1, 1.3\}$, while we vary $\alpha \in \{0.3, 6, \ldots, 12, 15\}$ and fix $\eta = 0.5$ for the heteroscedastic case. Within each combination of $\eta$ and $\rho$ (or $\alpha$ and $\rho$), we generate $X$, $A$, $B$ and $\Theta$ once and generate 100 times of the stochastic errors $W$ and $E$. 29
For each method with their estimator $\hat{\Theta}$ and the prediction $X \hat{F}$ (if available), we record the averaged Root Mean Square Error (RMSE) $\|\hat{\Theta} - \Theta\|_F / \sqrt{m}$ and the averaged Prediction Mean Square Error (PMSE) $\|X \hat{F} - XF\|_F^2 / (nm)$. We only report the results for $\rho = 0.5$ here as the ones for $\rho = 0$ are similar.

### 6.1.1 RMSE

The averaged RMSE of all methods are reported in Figure 2 for homoscedastic cases and Figure 3 for heteroscedastic cases. To illustrate the difference, we take the log$_{10}$ transformation.

**Homoscedastic cases:** HIVE dominates the other methods and has the closest performance to the Oracle across all settings. H-HIVE is the second best and has similar performance to HIVE when $p$ is small. This is expected since H-HIVE also works when the errors are homoscedastic. However, when $p$ is large, its performance deteriorates comparing to HIVE as $\eta$ increases such that the dense matrix $L$ has larger magnitude. The reason is that the condition $Rem(P_{B^*}) \leq c \sqrt{K}$ in Theorem 9 becomes restrictive for large $p$, small $m$ and large $\eta$ (say $\eta \geq 0.8$), since in this scenario the prediction error gets larger and so does $Rem(P_{B^*})$.

Among the competing methods, when $n > p$ (the first two panels of Figure 2), SVA also has good performance but is still outperformed by HIVE since SVA does not adapt to the sparsity structure of $\Theta^*$. OLS is comparable to Ridge. Lasso has clear advantage over Ridge when the signal is sparse enough, that is, when $\eta$ is small. HIVE-init outperforms both Lasso and Ridge. When $n < p$, SVA and OLS are not well defined and become infeasible in the third panel of Figure 2. HIVE-init has similar performance as Lasso but has larger error when $\eta$ increases. HIVE and H-HIVE dramatically reduce the error of the initial estimator HIVE-init in all setting. This agrees with the theoretical results in Remark 5.

![Figure 2: RMSE under the homoscedastic settings with $n = 100$.](image)

**Heteroscedastic cases:** Figure 3 shows that H-HIVE, tailored for the heteroscedastic error, has the smallest RMSE among all the methods and its advantage over the second best
method, HIVE, becomes evident when \( m \) is small (see the first and third panels) and the degree of heteroscedasticity is moderate or large (i.e., \( \alpha \geq 9 \)). This agrees with our theoretical analysis that the HIVE estimator may not be consistent when \( m \) is finite and the error is heteroscedastic. It is worth mentioning that when \( m \) is large (see the second panel), HIVE is nearly identical to H-HIVE suggesting similar performance between PCA and HetroPCA. This is expected in light of Remark 6. Finally, Ridge, Lasso and HIVE-init are robust to the degree of heteroscedasticity across all cases, whereas SVA shows inflated RMSE as the degree of heteroscedasticity (\( \alpha \)) increases when \( m \) is small (see the first panel).

6.1.2 PMSE

The PMSE for different methods are reported in Figure 4 for homoscedastic cases and in Figure 5 for heteroscedastic cases. Notice that OLS and SVA have the same PMSE and so are HIVE-init, HIVE and H-HIVE.

**Homoscedastic cases:** As seen in Figure 4, when \( n < p \) (the third panel), HIVE has much smaller PMSE than both Lasso and Ridge. This demonstrates the advantage of the proposed procedure in (2.3) for prediction. When \( p < n \) (the first two panels), HIVE and Lasso have comparable performance and clearly outperform OLS and Ridge for small \( \eta \) (i.e. the signal \( \Theta + L \) is approximately sparse). These findings are in line with Theorem 3 and its subsequent remarks.

**Heteroscedastic cases:** Figure 5 shows that all methods have robust prediction performance under the heteroscedastic cases and the advantage of HIVE (H-HIVE) becomes more evident when \( p > n \) (the last panel).
6.2 Performance of selecting $K$

We report our simulation results of selecting $K$ by using (5.1) (Ratio) and the permutation test (PA) in Buja and Eyuboglu (1992). In the setting of $n = 100$, $m = 150$, $p = 20$, both methods select $K$ consistently. This is expected for large $m$ and small $p$.

We mainly investigate the selection of $K$ in two settings: $n = 100$, $m = 20$, $p = 20$ and $n = 100$, $m = 20$, $p = 150$. For each setting, we fix $\mu_\Theta = \sigma_\Theta = 1$, $\eta = 0.3$, $\rho = 0.3$ and vary the signal-to-noise ratio (SNR) defined as $\lambda_K(B^T \Sigma_W B)/(m \tau^2)$, where $\tau^2 = 1$ and $B_{kj} \sim N(0,1)$.

We choose $\Sigma_W = \sigma_W^2 I_K$ with $\sigma_W \in \{0.1, 0.3, 0.5, \ldots, 1.3, 1.5\}$ such that SNR $\approx \sigma_W^2$. Recall that the true $K$ is equal to 3. Figure 6 shows the boxplot of the selected $K$ by using Ratio and PA over 100 simulations. It is clear that as long as the SNR is large enough, both methods consistently select $K$. By comparing the two panels, we can see that when $p$ is large, we need stronger SNR in order to consistently select $K$.

In practice, we recommend using PA when $m$ is small, say around 20. When $m$ is large or
moderate, PA becomes computationally expensive due to the implementation of SVD on the permuted data. For this reason, we recommend Ratio for moderate or large $m$.

![Boxplots of the selected $K$ using PA and Ratio in two settings.](image)

Figure 6: Boxplots of the selected $K$ using PA and Ratio in two settings.

## 7 Real data application

We apply our procedures, Algorithms 1 and 3, to two real word datasets: the Norwegian dataset and the yeast cross dataset. While prediction is not the main focus of our procedures, due to the lack of knowledge of the ground truth in real data application, we compare the performance of our procedures with several competing methods in terms of their prediction errors.

**Norwegian dataset.** This dataset available in Izenman (2008) was collected to study the effect of three variables $X_1$, $X_2$ and $X_3$ on the quality of the paper from a Norwegian paper factory. The quality of the paper is measured by 13 continuous responses while all $X_i$ taking values in $\{-1, 0, 1\}$ represent the location of the design point. In addition to the main effect terms ($X_1, X_2, X_3$), six second order interaction terms ($X_1^2, X_2^2, X_3^2, X_1X_2, X_1X_3, X_2X_3$) were also considered as predictors. In total, the dataset consists of $n = 29$ fully observed observations with $m = 13$ responses and $p = 9$ predictors. The design matrix is centered and standardized to unit variance while the response matrix is centered.

Bunea et al. (2011) showed that the data may exhibit a low-rank structure with estimated rank $K = 3$ via reduced-rank regression. This finding is consistent with Aldrin (1996) based on the smallest leave-one-out cross-validation (LOOCV) error, which is 326.2 (total sum of squared errors), over all possible ranks. A later analysis of Bunea et al. (2012) via the sparse reduced-rank regression (SSR) further reduces the LOOCV error to 304.5. Specifically, Bunea et al. (2012) estimated the coefficient matrix of the multivariate linear regression by

$$
\hat{F}_k = \min_{\text{rank}(F) \leq k} \|Y - X F\|_F^2 + 2\lambda\|F\|_{\ell_1/\ell_2}
$$

(7.1)
with $k = 3$ and $\lambda > 0$ selected from CV. The resulting estimator $\hat{F}_k$ is both low-rank and row-sparse, and selects 6 predictors by excluding the following three terms $X_1^2$, $X_1X_2$ and $X_2X_3$.

To compare the prediction performance, we applied HIVE in Algorithm 1 to this dataset and the permutation test in Section 5.1 for estimating $K$ as $m$ is small. Our procedure yields $\hat{K} = 3$. The results from the H-HIVE algorithm are similar and thus omitted. For comparison, we also applied Ridge, group-lasso (Lasso) and SVA to this dataset (note that the prediction of SVA is the same as OLS). The LOOCV errors for all methods are summarized in Table 1. The HIVE algorithm has the smallest LOOCV error among all methods. Thus, our approach yields the most accurate prediction.

| Method     | HIVE | Ridge | Lasso | SVA  | RRR  | SRR  |
|------------|------|-------|-------|------|------|------|
| LOOCV error| 288.9| 324.3 | 317.3 | 338.1| 326.2| 304.5|

In addition, the results in Table 1 imply that the low-rank structure of the coefficient matrix in RRR and SRR may not be sufficient to model the association between the predictors and responses. Bunea et al. (2011) showed that the reduced-rank regression by using all $p$ predictors can explain 86.9% of the total variation of $Y$ quantified by $\text{tr}(Y^T X (X^T X)^{-1} X^T Y)$ (see Izenman (2008) for the definition). Note that we can rewrite model (1.4) as a reduced-rank regression of $Y - X\Theta$ on $X$. By replacing $\Theta$ with our estimator $\hat{\Theta}$, we can show that our model (1.4) can explain 98.6% of the total variation, a much higher percentage than the reduced-rank regression. This implies that our model may provide a better fit to the data than the reduced-rank regression. The main reason is that model (1.4) is able to capture the sparse signal that cannot be explained by the low rank structure. To quantify this statement, we calculate, in Table 2, the $\ell_2$ norm of rows of our estimator $\hat{L} + \hat{\Theta}$ corresponding to all the predictors. As a comparison, we also compute the reduced-rank estimator $\tilde{L}$ by regressing $Y$ on $X$ directly. The results are also shown in Table 2. Similar to the results from the SRR, the estimator $\hat{L}$ corresponding to the three predictors $X_1^2$, $X_1X_2$ and $X_2X_3$ has small $\ell_2$ norm. However, the association between the three predictors and responses becomes evidently stronger by using our estimator $\hat{\Theta} + \hat{L}$, as the sparse signal $\hat{\Theta}$ is taken into account. This suggests that model (1.1) can successfully capture both the low rank signal and the sparse signal, whereas the latter is omitted in the (sparse) reduced-rank regression.

|        | $X_1$ | $X_2$ | $X_3$ | $X_1^2$ | $X_2^2$ | $X_3^2$ | $X_1X_2$ | $X_1X_3$ | $X_2X_3$ |
|--------|-------|-------|-------|---------|---------|---------|---------|---------|---------|
| $\tilde{L}$ | 1.33  | 0.60  | 1.05  | **0.28**| 0.44    | 0.64    | **0.14**| 0.71    | **0.35**|
| $\hat{\Theta} + \hat{L}$ | 1.61  | 0.94  | 1.15  | **0.59**| 0.56    | 0.78    | **0.23**| 0.88    | **0.54**|

Table 2: $\ell_2$ norms of rows of $\tilde{L}$ and $\hat{\Theta} + \hat{L}$. The bold numbers correspond to the three excluded predictors, $X_1^2$, $X_1X_2$ and $X_2X_3$ in Bunea et al. (2012).
**Yeast cross dataset.** The yeast cross dataset consists of 1,008 prototrophic haploid segregants from a cross between a laboratory strain and a wine strain of yeast. This dataset was collected via high-coverage sequencing and consists of genotypes at 30,594 high-confidence single-nucleotide polymorphisms (SNPs) that distinguish the strains and densely cover the genome. There are 46 traits in this dataset corresponding to the measured growth under multiple conditions, including different temperatures, pHs and carbon sources, as well as addition of metal ions and small molecules (Bloom et al., 2013). The goal is to study the relationship between genotypes and traits, which could be used for predicting traits or selecting significant genotypes for further scientific investigation (Bloom et al., 2013). A multivariate linear regression by regressing traits on genotypes could be suitable for this purpose. However, it is likely that there exist hidden factors that also affect traits. We thus fit our model (1.4) for prediction and variable selection. After removing the segregants with missing values in traits and SNPs which have Pearson correlations above 0.97, we end up with \( n = 303 \) segregants with \( m = 46 \) traits and \( p = 571 \) SNPs.

To evaluate the prediction performance, we randomly split the data into 70% training set and 30% test set. We centered and normalized the SNPs in the training set to zero mean and unit variance. Traits in the training set were also centered. The corresponding means and scales from the training set were used to standardize the test set. We then applied the HIVE Algorithm 1 together with group-lasso (Lasso) and Ridge to the training set and evaluated the fitted model on the test set. The test mean square errors (MSE) of Lasso and Ridge are 7.29 and 6.29, respectively, while HIVE has a smaller test MSE 5.92. This suggests that HIVE has better prediction performance than Lasso and Ridge. We then refitted the model to the whole dataset and applied HIVE, H-HIVE and Lasso for variable selection. Lasso and HIVE select, respectively, 261 and 259 SNPs with 205 common ones. For H-HIVE, it selects 263 SNPs in which 222 SNPs are identical to those selected from Lasso. The difference of the selected SNPs between HIVE (H-HIVE) and Lasso is due to the fact that Lasso does not account for the potential hidden variables. We expect that the results from HIVE (H-HIVE) may provide new insight on understanding how SNPs are associated with different traits. For instance, further confirmatory analysis such as controlled experiments can be conducted by the investigators to study the effect of the selected SNPs.

8 Discussion

In this paper, we study the high-dimensional multivariate regression model with hidden variables. We establish sufficient and necessary conditions for model identifiability. We propose the HIVE algorithm for estimating the sparse coefficient matrix \( \Theta^* \), which is adaptive to the unknown sparsity of \( \Theta^* \). The algorithm is further extended to the setting with heteroscedastic noise. Theoretically, we establish non-asymptotic upper bounds for the errors of our estimator, which are valid for any finite \( n, p, m \) and \( K \).

There are several future directions that are worthy of further investigation. First, we plan

---

3The dataset is downloaded from [http://genomics-pubs.princeton.edu/YeastCross_BYxRM/home.shtml](http://genomics-pubs.princeton.edu/YeastCross_BYxRM/home.shtml)
to study the variable selection property of the proposed algorithm. In this paper, we focus on
the adaptive estimation of the coefficient matrix. To establish the variable selection consistency
property, a different set of conditions (e.g., minimum signal strength condition) are required.
We aim to investigate this property in some future work. Second, it is of great interest to study
how to construct confidence intervals or hypothesis tests for the high-dimensional matrix $\Theta^*$. The inference results can be further used to control the false discovery rate (FDR) in multiple
testing, which is of central importance in many biological applications. We refer to the SVA
literature for discussions on the FDR control. Third, the proposed estimation procedures can
be extended to handle a variety of different sparsity patterns of $\Theta^*$. Since we assume that there
exists a small subset of common features associated with the responses, we apply the group-
lasso penalty to estimate the row-sparse matrix $\Theta^*$. In other applications, a different sparsity
structure of $\Theta^*$ may be more suitable (for instance, column-sparsity or block-sparsity). One
can modify our procedure to account for the specific sparsity structures.

Acknowledgement

Yang Ning was partially supported by NSF-DMS 1854637. Xin Bing was supported in part by
NSF-DMS 1712709

References

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

Aldrin, M. (1996). Moderate projection pursuit regression for multivariate response data.
Computational Statistics & Data Analysis 21 501 – 531.
URL http://www.sciencedirect.com/science/article/pii/0167947394000298

Anderson, T. W. (1984). An introduction to multivariate statistical analysis. Wiley Series in
Probability and Statistics, Wiley.

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. (2008). Forecasting economic time series using targeted predictors. Journal
of Econometrics 146 304 – 317. Honoring the research contributions of Charles R. Nelson.

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large dimensional
sample covariance matrix. Ann. Probab. 21 1275–1294.

Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root
lasso in nonparametric regression. Ann. Statist. 42 757–788.
URL https://doi.org/10.1214/14-AOS1204

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and
Bing, X., Bunea, F., Wegkamp, M. and Strimas-Mackey, S. (2019). Essential regression. 
*Ann. Statist.* **37** 1705–1732. 
URL https://doi.org/10.1214/08-AOS620

Bing, X. and Wegkamp, M. H. (2019). Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.* **47** 3157–3184. 
URL https://doi.org/10.1214/18-AOS1774

Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature* **494** 234–237.

Bühlmann, P. and Van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research* **27** 509–540. PMID: 26811132. 
URL https://doi.org/10.1207/s15327906mbr2704_2

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. 
URL https://doi.org/10.1214/11-AOS876

Bunea, F., She, Y. and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.* **40** 2359–2388. 
URL https://doi.org/10.1214/12-AOS1039

Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *J. ACM* **58** 11:1–11:37. 
URL http://doi.acm.org/10.1145/1970392.1970395

Candès, E. J. and Tao, T. (2009). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080.

Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory* **57** 2342–2359.

Čevid, D., Bühlmann, P. and Meinshausen, N. (2018). Spectral deconfounding via perturbed sparse linear models. *arXiv preprint arXiv:1811.05352* .

Chakraborty, S., Datta, S. and Datta, S. (2012). Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics* **28** 799–806. 
URL https://doi.org/10.1093/bioinformatics/bts022

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

Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics* 1935–1967.

Chernozhukov, V., Hansen, C. and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.* **45** 39–76. 
URL https://doi.org/10.1214/16-AOS1434

Chernozhukov, V. and Zhou, H. (2015). Inference in high-dimensional models. In *High-Dimensional Econometrics: A Non-Asymptotic Approach* (pages 1–74). Cambridge University Press.
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.

Diaz, E. (2017). Causality and surrogate variable analysis. *arXiv preprint arXiv:1704.00588*.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37.

Dragomir, S. S. (2015). Reverses of schwarz inequality in inner product spaces with applications. *Mathematische Nachrichten* **288** 730–742.

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.

Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.

Giraud, C. (2011). Low rank multivariate regression. *Electron. J. Statist.* **5** 775–799.

Giraud, C. (2015). *Introduction to high-dimensional statistics*. Monographs on Statistics and Applied Probability.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K. and Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13** 86.

Hox, J. J. and Bechger, T. M. (1998). An introduction to structural equation modeling .

Hsu, D., Kakade, S. M. and Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory* **57** 7221–7234.

Hsu, D., Kakade, S. M. and Zhang, T. (2014). Random design analysis of ridge regression. *Found. Comput. Math.* **14** 569–600.

Izenman, A. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* **5** 248–262.

Izenman, A. (2008). *Modern Multivariate. Statistical Techniques: Regression, Classification and Manifold Learning*. Springer.

Koltchinskii, V., Louini, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329.

URL: https://doi.org/10.1214/11-AOS894
Lam, C. and Yao, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* 40 694–726.

Lee, S., Sun, W., Wright, F. A. and Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* 104 303–316.

Lee, S., Sun, W., Wright, F. A. and Zou, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* 104 303–316.

URL https://doi.org/10.1093/biomet/asx018

Lee, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* 3 1724–1735.

Lee, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* 105 18718–18723.

URL https://www.pnas.org/content/105/48/18718

Lounici, K., Pontil, M., Van de Geer, S. and Tsybakov, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* 39 2164–2204.

URL https://doi.org/10.1214/11-AOS986

Mckennan, C. and Nicolae, D. (2019). Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika* 106 823–840.

URL https://doi.org/10.1093/biomet/asz037

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics, Wiley.

Negahban, S. and Wainwright, M. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* 39 1069–1097.

Obozinski, G., Wainwright, M. and Jordan, M. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* 39 1–47.

Reinsel, G. and Velu, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer.

Rohde, A. and Tsybakov, A. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* 39 887–930.

Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* 59 3434–3447.

Srivastava, M. and Khatri, C. (1979). *An introduction to multivariate statistics*. North-Holland/New York.

Sun, Y., Zhang, N. R. and Owen, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *Ann. Appl. Stat.* 6 1664–1688.

URL https://doi.org/10.1214/12-AOAS561

Teschendorff, A. E., Zhuang, J. and Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 27 1496–1505.

URL https://doi.org/10.1093/bioinformatics/btr171

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*. Cambridge University Press, 210268.
Wang, J., Zhao, Q., Hastie, T. and Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* 45 1863–1894.
URL https://doi.org/10.1214/16-AOS1511

Yu, Y., Wang, T. and Samworth, R. J. (2014). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* 102 315–323.
URL https://doi.org/10.1093/biomet/asv008

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69 329–346.
URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2007.00591.x

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* 68 49–67.

Zhang, A., Cai, T. T. and Wu, Y. (2018). Heteroskedastic pca: Algorithm, optimality, and applications.
A A variant of the robust \( \sin \Theta \) theorem in Zhang et al. (2018)

We state a variant of Theorem 3 (Robust \( \sin \Theta \) theorem) in Zhang et al. (2018). Let \( N, M \) and \( Z \) be \( m \times m \) deterministic symmetric matrices satisfying

\[
N = M + Z
\]

(A.1)

where \( N \) is the observation matrix while \( M \) is the matrix of interest with \( \text{rank}(M) = K \). Let \( V \in \mathbb{R}^{m \times K} \) denote the first \( K \) eigenvectors of \( M \) with non-increasing eigenvalues denoted as \( \lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_K(M) \). Further let \( \hat{V} \) be the output of Algorithm 2 applied to \( N \). Define \( \sin \Theta(\hat{V}, V) := \hat{V}^T V \) where \( \hat{V}_\perp \in \mathbb{R}^{m \times (m-K)} \) is the orthogonal complement of \( \hat{V} \) such that \([\hat{V}, \hat{V}_\perp]\) is the complete orthogonal matrix. The following theorem provides upper bounds for the Frobenius norm of \( \sin \Theta(\hat{V}, V) \). For notational simplicity, for any matrix \( D \), let \( G(D) \) be the matrix with diagonal entries equal to those of \( D \) and off-diagonal entries equal to zero. Let \( \Gamma(D) = D - G(D) \).

**Theorem 12.** Suppose \( M \in \mathbb{R}^{m \times m} \) is a rank-\( K \) symmetric matrix and \( V \) consists its first \( K \) eigenvectors. Assume there exists a universal constant \( c > 0 \) such that

\[
\max_{1 \leq j \leq m} \| e_j^T V \|_2^2 \cdot \frac{\lambda_1(M)}{\lambda_K(M)} \leq c
\]

(A.2)

is satisfied and \( \| \Gamma(Z) \|_F \leq c\sqrt{K}\lambda_K(M) \). Then the output \( \hat{V} \) of HeteroPCA(\( N, K, T \)) in Algorithm 2 with \( T = \Omega(1 \log \frac{\sqrt{K}\lambda_K(M)}{\| \Gamma(Z) \|_F}) \) satisfies

\[
\| \sin \Theta(\hat{V}, V) \|_F \lesssim \frac{\| \Gamma(Z) \|_F}{\lambda_K(M)} \wedge \sqrt{K}.
\]

**Proof of Theorem 12.** The proof follows the same arguments as that of Theorem 7 (General robust \( \sin \Theta \) theorem) in Zhang et al. (2018) by taking the sparsity set \( \mathcal{G} = \{(i, i) : 1 \leq i \leq m\} \) and \( b = \eta = 1 \). We will use the same notations as Zhang et al. (2018).

Define \( T_0 = \| \Gamma(N - M) \|_F = \| \Gamma(Z) \|_F \) and \( K_t = \| N^{(t)} - M \|_F \) for \( t = 0, 1, \ldots \). Note that \( \| H \|_F \leq \| \Gamma(H) \|_F + \| G(H) \|_F \) for all matrix \( H \in \mathbb{R}^{m \times m} \). We then revisit the three steps in the proof of Theorem 7 in Zhang et al. (2018) by only restating the main results and differences from Zhang et al. (2018). We use \( V \) and \( V^{(t)} \) in lieu of \( U \) and \( U^{(t)} \) in the original proofs of Zhang et al. (2018) for all \( t \geq 0 \).

Step 1: The initial errors satisfy

\[
K_0 \leq \| \Gamma(Z) \|_F + \| G(P_Y M P_V) \|_F \\
\leq \| \Gamma(Z) \|_F + \| M \|_F \max_i \| e_i^T V \|_2 \\
\leq \| \Gamma(Z) \|_F + c\sqrt{K}\lambda_K(M)
\]

for some sufficiently small constant \( c > 0 \) where the first inequality follows from the same arguments in Zhang et al. (2018), the second inequality uses the modified Lemma 3 and the third inequality uses \( \| M \|_F \leq \sqrt{K}\| M \|_{op} = \sqrt{K}\lambda_1(M) \) together with condition (A.2).
Step 2: From the proof of Zhang et al. (2018), one needs to upper bound the two terms on the right hand side of the following display

\[ K_t \leq \| \Gamma(N^{(t)}) - M \|_F + \| G(N^{(t)}) - M \|_F. \]

The first term satisfies

\[ \| \Gamma(N^{(t)}) - M \|_F = \| \Gamma(N - M) \|_F = T_0 \quad \text{(A.3)} \]

for all \( t \geq 0 \). The second term can be upper bounded by (see display (6.15) in Zhang et al. (2018))

\[ \| G(N^{(t)}) - M \|_F \leq \| G(P_V(N^{(t-1)} - M)) \|_F + \| (P_{V^{(t-1)}} - P_V) (N^{(t-1)} - M) \|_F \]

+ \( \| G(P_{V^{(t-1)}}^T M) \|_F. \]

To bound them separately, we have

\[ \| G(P_V(N^{(t-1)} - M)) \|_F \leq \| N^{(t-1)} - M \|_F \max_i \| e_i^T V \|_2 = K_{t-1} \max_i \| e_i^T V \|_2 \quad \text{(A.4)} \]

by Lemma 3, and

\[ \| G(P_{V^{(t-1)}}^T M) \|_F \leq \| G(P_{V^{(t-1)}}^T M P_V) \|_F \]

\[ \leq \| P_{V^{(t-1)}}^T M \|_F \max_i \| e_i^T V \|_2 \]

\[ \leq 2 \| N^{(t-1)} - M \|_F \max_i \| e_i^T V \|_2 \]

\[ = 2K_{t-1} \max_i \| e_i^T V \|_2 \quad \text{(A.5)} \]

by using Lemma 3 in the second inequality and using Lemma 7 in Zhang et al. (2018) in the third inequality. Finally, the proof of Theorem 7 in Zhang et al. (2018) shows that

\[ \| P_{V^{(t-1)}} - P_V \|_{op} \leq \frac{4 \| N^{(t-1)} - M \|_{op}}{\lambda_K(M)} \land 1 \leq \frac{4K_{t-1}}{\lambda_K(M)} \]

which further gives

\[ \| (P_{V^{(t-1)}} - P_V) (N^{(t-1)} - M) \|_F \leq \| P_{V^{(t-1)}} - P_V \|_{op} \| N^{(t-1)} - M \|_F \leq \frac{4K_{t-1}}{\lambda_K(M)} \quad \text{(A.6)} \]

Collecting (A.3) – (A.6) yields

\[ K_t \leq T_0 + 3K_{t-1} \max_i \| e_i^T V \|_2 + \frac{4K_{t-1}^2}{\lambda_K(M)} \]

Step 3: Finally, by using the same induction arguments in Zhang et al. (2018), under (A.2) and \( \| \Gamma(Z) \|_F = T_0 \leq c\sqrt{K} \lambda_K(M) \) with sufficiently small \( c \), one can show that for all \( t \geq 0 \),

\[ K_t \leq 2T_0 + \frac{\sqrt{K} \lambda_K(M)}{2t+4}. \]

Therefore, for all \( t \geq \Omega(1 \lor \log(\sqrt{K} \lambda_K(M)/T_0)) \), one has \( K_t \leq 3T_0 \). The proof is completed by invoking the variant of Davis-Kahan’s \( \sin \Theta \) theorem (Yu et al., 2014) applied to \( N^{(t)} \) and \( M \) with \( d = s = K, r = 1 \) and \( \lambda_{K+1}(M) = 0 \). \( \square \)
The following lemma is a variant of Lemma 1 in Zhang et al. (2018) and it provides upper bounds for the Frobenius norms of some diagonal matrices. Recall that $G(A)$ has the same diagonal elements with $A$ but all off-diagonal entries equal to zero.

**Lemma 3.** For any two orthogonal matrices $U, V \in \mathbb{R}^{m \times K}$, let $P_U = UU^T$ and $P_V = VV^T$. Then for any matrix $A \in \mathbb{R}^{m \times m}$, we have

\[
\|G(P_U A)\|_F^2 \leq \|A\|_F^2 \max_{1 \leq i \leq m} \|e_i^T U\|_2^2,
\]

\[
\|G(AP_V)\|_F^2 \leq \|A\|_F^2 \max_{1 \leq j \leq m} \|e_j^T V\|_2^2,
\]

\[
\|G(P_U A P_V)\|_F^2 \leq \|A\|_F^2 \left( \max_{1 \leq i \leq m} \|e_i^T U\|_2^2 \wedge \max_{1 \leq j \leq m} \|e_j^T V\|_2^2 \right).
\]

**Proof.** Since $G(A)$ has non-zero entries only on the diagonal, by writing $A = (A_1, \ldots, A_m)$, one has

\[
\|G(P_U A)\|_F^2 = \sum_{i=1}^m (e_i^T U U^T A_i)^2 \leq \max_{1 \leq i \leq m} \|e_i^T U\|_2^2 \sum_{i=1}^m A_i^T U U^T A_i
\]

\[
= \max_{1 \leq i \leq m} \|e_i^T U\|_2^2 \cdot \text{tr}(A^T U U^T A)
\]

\[
\leq \max_{1 \leq i \leq m} \|e_i^T U\|_2^2 \cdot \|A\|_F^2
\]

where we used Cauchy-Schwarz inequality in the first line and the inequality $\text{tr}(A^T U U^T A) \leq \|A\|_F^2 \|U U^T\|_{op} \leq \|A\|_F^2$ in the last line. Since $G(P_U A)$ is symmetric to $G(AP_V)$, the second result follows. The last result follows by using the first two results together with $\|P_V\|_{op} \leq 1$ and $\|P_U\|_{op} \leq 1$. \hfill \Box

## B Main proofs

### B.1 Proof of Proposition 1: the necessity of Assumption 1 and rank$(\Sigma_W) = K$

We will show that if Assumption 1 does not hold then $\Theta$ is not identifiable. For simplicity, we suppress the super script $*$ for related parameters. We will use lower case $x, y, z$ to denote the realizations of random variables $X, Y, Z$. Since the joint density of $(X, Y)$ can be factorized as

\[
f_{(Y,X)}(y,x) = f_{Y|X}(y|x)f_X(x) = f_X(x) \cdot \int f_{Y|X}(y|x,z)f_{Z|X}(z|x)dz,
\]

under model (1.1), we can write the likelihood of $(Y, X)$ as

\[
f_{(Y,X)}(y,x; \Theta, B, \mu_E, \nu_{Z,X}) = f_X(x) \cdot \int f_E\left(y - \Theta^T x - B^T z; \mu_E\right) f_{Z|X}(z; \nu_{Z,X})dz \quad (B.1)
\]

where $f_E$ denotes the density of $E$ parametrized by $\mu$ and $f_{Z|X}$ is the conditional p.d.f. of $Z = z|X = x$ parametrized by $\nu_{z,x}$. We emphasize that $\nu_{z,x}$ depends on $X = x$ as $Z$ is correlated with $X$. Write

\[
\tilde{\Theta} = \Theta - \Theta P_B = \Theta - \Theta B^T (BB^T)^{-1} B, \quad \tilde{Z} = Z + (BB^T)^{-1} B \Theta^T X.
\]
Note that $\tilde{\Theta} \neq \Theta$ when $\Theta P_B \neq 0$. We will prove the following to conclude part (1),

$$f(Y,X)(y,x; \Theta, B, \mu, \nu_{z,x}) = f(Y,X)(\tilde{y}, \tilde{x}; \tilde{\Theta}, B, \mu, \nu_{\tilde{z}, \tilde{x}}).$$

Write $\Delta = \Theta B^T(BB^T)^{-1}$ such that $\tilde{Z} = Z + \Delta^T X$. Observe that the density functions of $Z|X$ and $\tilde{Z}|X$ only differ by a shift of the mean. We then deduce

$$f_{\tilde{Z}|X}(z; \nu_{\tilde{z}, x}) = f_{Z|X}(z - \Delta^T x; \nu_{z, x}),$$

from which we further obtain

$$f_{(Y,X)}(y, x; \tilde{\Theta}, B, \mu, \nu_{\tilde{z}, x})
= f_X(x) \cdot \int f_E \left( y - \tilde{\Theta}^T x - B^T z; \mu \right) f_{\tilde{Z}|X}(z; \nu_{\tilde{z}, x}) \, dz
= f_X(x) \cdot \int f_E \left( y - \tilde{\Theta}^T x - B^T z; \mu \right) f_{Z|X}(z - \Delta^T x; \nu_{z, x}) \, dz
= f_X(x) \cdot \int f_E \left( y - \tilde{\Theta}^T x - B^T t - B^T \Delta^T x; \mu \right) f_{Z|X}(t; \nu_{z, x}) \, dt
= f_X(x) \cdot \int f_E \left( y - \Theta^T x - B^T t; \mu \right) f_{Z|X}(t; \nu_{z, x}) \, dt
= f_{(Y,X)}(y, x; \Theta, B, \mu, \nu_{z, x}).$$

We use (B.1) in the first and last equality, set $t = z - \Delta^T x$ to derive the third equality and use

$$\tilde{\Theta}^T x = \Theta^T x - B^T (BB^T)^{-1} B \Theta^T x = \Theta^T x - B^T \Delta^T x$$

to arrive at the fourth equality. This completes the proof of part (1).

To show part (2), under model (1.4)

$$Y = (\Theta + DB)^T X + B^T W + E,$$

suppose $E[W|X] = 0$ and Assumption 1 holds, that is, $\Theta P_B = 0$, or equivalently, $\Theta B^T = 0$. It remains to show that $\Theta^*$ is not identifiable if $\operatorname{rank}(\Sigma_W) < K$.

Suppose $\operatorname{rank}(\Sigma_W) < K$. Then there exists a subspace $S \subseteq \mathbb{R}^K$ with $\operatorname{rank}(S) = K - \operatorname{rank}(\Sigma_W)$ such that

$$v^T \Sigma_W v = 0, \quad \text{for any } v \in S. \tag{B.2}$$

Let $P_S \in \mathbb{R}^{K \times K}$ denote the projection matrix onto $S$ and $P_S^\perp = I_K - P_S$. We will first show that $\operatorname{rank}(\Sigma_W) < K$ implies $DP_S \neq 0$, and then prove that $\Theta^*$ is not identifiable if $DP_S \neq 0$.

We prove $DP_S \neq 0$ by contradiction. Suppose $DP_S = 0$ such that $D = DP_S^\perp$. This implies that the row space of $D$ lies in $S^\perp$ which is also the row space of $\Sigma_W$. By recalling that

$$\operatorname{Cov}(Z) = D^T \operatorname{Cov}(X) D + \Sigma_W$$

and $\operatorname{rank}(\operatorname{Cov}(Z)) = K$, we have

$$K = \operatorname{rank}(\operatorname{Cov}(Z)) \leq \operatorname{rank}(\Sigma_W).$$
This contradicts with rank(ΣW) < K hence proves DPₘ ≠ 0.

We proceed to show that Θ∗ is not identifiable given that DPₘ ≠ 0. Write

\[ \tilde{B} = P_B^\perp B = B - P_B B, \quad \tilde{\Theta} = \Theta + DP_B B. \]

such that \( \tilde{\Theta} \neq \Theta \) as DPₘ ≠ 0. Consider model

\[ Y = (\tilde{\Theta} + D \tilde{B})^T X + \tilde{B}^T W + E. \]

Note that (B.2) implies \( \tilde{B}^T W = B^T W \) a.s. and \( \tilde{\Theta} + D \tilde{B} = \Theta + DB \). Since, by using \( \Theta B^T = 0 \),

\[ \tilde{\Theta} \tilde{B}^T = \Theta B^T P_B^\perp + DPP_B^\perp = 0, \]

this implies \( \tilde{\Theta} P_B^\perp = 0 \), hence Assumption 1 still holds for \( \tilde{\Theta} \) and \( \tilde{B} \). We conclude that \( \Theta \) is not identifiable and the proof is then complete.

\[ \blacksquare \]

B.2 Proofs of Propositions 2 and 8: identifiability for both homoscedastic and heteroscedastic cases

\textit{Proof of Proposition 2.} Model (1.4) implies

\[ F^* = \Theta^* + A^* B^* = [\text{Cov}(X)]^{-1} \text{Cov}(X,Y), \]

from which we can identify

\[ \Sigma_e := \text{Cov}(\varepsilon) = \text{Cov}(Y - (F^*)^T X). \]

By further using \( \Sigma_E = \tau^2 I_m \), we have

\[ \Sigma_e = (B^*)^T \Sigma_W B^* + \Sigma_E = (B^*)^T \Sigma_W B^* + \tau^2 I_m \]

Recall that \( P_{B^*} \) denotes the projection onto the row space of \( B^* \). One can identify it from

\[ P_{B^*} = U_K U_K^T \]

where \( U_K \) are the eigenvectors of \( \Sigma_e \) corresponding to the largest \( K \) eigenvalues. Projecting \( Y \) onto \( P_{B^*}^\perp = I_m - P_{B^*} \) gives

\[ P_{B^*}^\perp Y = (\Theta^* P_{B^*}^\perp)^T X + P_{B^*}^\perp E = (\Theta^*)^T X + P_{B^*}^\perp E \]

where we used Assumption 1 to arrive at the second equality. We then conclude

\[ \Theta^* = [\text{Cov}(X)]^{-1} \text{Cov}(X, P_{B^*}^\perp Y) \]

which is uniquely defined.

\[ \blacksquare \]

\textit{Proof of Proposition 8.} The proof of Proposition 8 uses the same arguments as that of Proposition 2 except the identifiability of \( P_{B^*} \), the projection onto the row space of \( B^* \). After identifying \( \Sigma_e \), note that

\[ \Sigma_e = (B^*)^T \Sigma_W B^* + \text{diag}(\tau_1^2, \ldots, \tau_m^2). \]

Under (4.2), applying Theorem 12 with \( N = \Sigma_e, M = (B^*)^T \Sigma_W B^*, Z = \text{diag}(\tau_1^2, \ldots, \tau_m^2) \) and \( T \to \infty \) identifies \( P_{B^*} \). Hence the identifiability of \( \Theta^* \) follows from the same arguments in the previous proof.

\[ \blacksquare \]
B.3 Proof of Theorem 3: in-sample prediction risk

Before proving Theorem 3, we first prove Lemma 1 as it provides a simpler analytical expression for the later proof.

Proof of Lemma 1. From (2.3), for any fixed \( \Theta \), we have

\[
\hat{L}(\Theta) = \arg \min_L \frac{1}{n} \|Y - X\Theta - XL\|_F^2 + \lambda_2 t_2 \|L\|_F^2 = (X^TX + n\lambda_2 I_p)^{-1}X^T(Y - X\Theta).
\]

Plugging this into (2.3) yields

\[
\hat{\Theta} = \arg \min_\Theta \frac{1}{n} \|Y - P_{\lambda_2} Y + P_{\lambda_2} X\Theta - X\Theta\|_F^2 + \lambda_1 \|\Theta\|_{\ell_1/\ell_2} + \lambda_2 \|\hat{L}(\Theta)\|_F^2
\]

\[
= \arg \min_\Theta \frac{1}{n} \|Q_{\lambda_2}(Y - X\Theta)\|_F^2 + \lambda_1 \|\Theta\|_{\ell_1/\ell_2} + \lambda_2 \|(X^TX + n\lambda_2 I_p)^{-1}X^T(Y - X\Theta)\|_F^2.
\]

Since

\[
\frac{1}{n} \|Q_{\lambda_2}(Y - X\Theta)\|_F^2 + \lambda_2 \|(X^TX + n\lambda_2 I_p)^{-1}X^T(Y - X\Theta)\|_F^2 = \text{tr} \left\{ (Y - X\Theta)^T \left[ \frac{1}{n}Q_{\lambda_2}^2 + \lambda_2 X(X^TX + n\lambda_2 I_p)^{-2}X^T \right] (Y - X\Theta) \right\}
\]

and

\[
\frac{1}{n}Q_{\lambda_2}^2 + \lambda_2 X(X^TX + n\lambda_2 I_p)^{-2}X^T
\]

\[
= \frac{1}{n} \left[ I_p + X(X^TX + n\lambda_2 I_p)^{-1}X^TX(X^TX + n\lambda_2 I_p)^{-1}X^T - 2X(X^TX + n\lambda_2 I_p)^{-1}X^T + n\lambda_2 X(X^TX + n\lambda_2 I_p)^{-2}X^T \right]
\]

\[
= \frac{1}{n} \left[ I_p - X(X^TX + n\lambda_2 I_p)^{-1}X^T \right]
\]

\[
= \frac{1}{n}Q_{\lambda_2}
\]

by using \( n\lambda_2 X(X^TX + n\lambda_2 I_p)^{-2}X^T = X(X^TX + n\lambda_2 I_p)^{-1}(n\lambda_2 I_p)(X^TX + n\lambda_2 I_p)^{-1}X^T \) in the last line, we obtain

\[
\hat{\Theta} = \arg \min_\Theta \frac{1}{n} \|Q_{\lambda_2}^{1/2}(Y - X\Theta)\|_F^2 + \lambda_1 \|\Theta\|_{\ell_1/\ell_2}.
\]

This proves (2.5). As a result, we have

\[
X\hat{L}(\hat{\Theta}) = X(X^TX + n\lambda_2 I_p)^{-1}X^T(Y - X\hat{\Theta}) = P_{\lambda_2}Y - P_{\lambda_2}X\hat{\Theta}
\]

hence

\[
X\hat{F} = X\hat{L}(\hat{\Theta}) + X\hat{\Theta} = P_{\lambda_2}Y + Q_{\lambda_2}X\hat{\Theta}.
\]

The proof is complete. □
To prove Theorem 3, recall that, for any \((\Theta_0, L_0)\) such that \(F^* = \Theta_0 + L_0\),

\[
\begin{align*}
X \hat{F} - XF^* &= X \hat{L} - XL_0 + X \hat{\Theta} - X \Theta_0 \\
&= P_{\lambda_2}(Y - X \Theta_0) - XL_0 + Q_{\lambda_2}(X \hat{\Theta} - X \Theta_0). \quad \text{(B.3)}
\end{align*}
\]

The result of Theorem 3 follows by invoking Lemmas 4–5 and noting that

\[
\frac{1}{n} \|X \hat{F} - XF^*\|_F^2 \leq \frac{2}{n} \|P_{\lambda_2}(Y - X \Theta_0) - XL_0\|_F^2 + \frac{2}{n} \|Q_{\lambda_2}(X \hat{\Theta} - X \Theta_0)\|_F^2
\]

from the basic inequality \((a + b)^2 \leq 2(a^2 + b^2)\). We then proceed to upper bound the two terms on the right hand side separately.

**Lemma 4.** Under conditions in Theorem 3, with probability \(1 - \epsilon\),

\[
\frac{1}{n} \|P_{\lambda_2}(Y - X \Theta_0) - XL_0\|_F^2 \leq 2 \|Q_{\lambda_2}\|_{op} \cdot \lambda_2 \text{tr} \left[ L_0^T \hat{\Sigma} (\hat{\Sigma} + \lambda_2 I_p)^{-1} L_0 \right] + \frac{2V_2}{n} \left[ \text{tr}(P_{\lambda_2}^2) + \sqrt{2 \|P_{\lambda_2}\|_{op} \log(m/\epsilon)} \right]^2
\]

where \(V_\epsilon\) is defined in (3.3).

**Proof.** By \(Y = X \Theta_0 + XL_0 + \epsilon\) and the basic inequality \((a + b)^2 \leq 2(a^2 + b^2)\), we have

\[
\|P_{\lambda_2}(Y - X \Theta_0) - XL_0\|_F^2 \leq 2 \|P_{\lambda_2} \|_F^2 + 2 \|Q_{\lambda_2} XL_0\|_F^2.
\]

Note that the second term satisfies

\[
\|Q_{\lambda_2} XL_0\|_F^2 \leq \|Q_{\lambda_2}\|_{op} \|Q_{\lambda_2}^{1/2} XL_0\|_F^2 = \|Q_{\lambda_2}\|_{op} \cdot n \lambda_2 \text{tr} \left[ L_0^T \hat{\Sigma} (\hat{\Sigma} + \lambda_2 I_p)^{-1} L_0 \right] \quad \text{(B.5)}
\]

by using Fact 1 in the second equality. The result follows by invoking Lemma 12 for the term \(\|P_{\lambda_2} \|_F^2\).

**Lemma 5.** Under conditions in Theorem 3, with probability \(1 - \epsilon'\),

\[
\frac{1}{n} \|Q_{\lambda_2} X (\hat{\Theta} - \Theta_0)\|_F^2 \leq 4 \|Q_{\lambda_2}\|_{op} \cdot \max \left\{ 4 \lambda_2 \text{tr} \left[ L_0^T \hat{\Sigma} (\hat{\Sigma} + \lambda_2 I_p)^{-1} L_0 \right], \frac{\lambda_2^2}{\kappa_1(1/2, \Theta_0, \lambda_1, \lambda_2)^2} \right\}
\]

**Proof.** Write \(\tilde{Y} = Q_{\lambda_2}^{1/2} Y\) and \(\tilde{X} = Q_{\lambda_2}^{1/2} X\). Starting with (2.5), we have

\[
\frac{1}{n} \|\tilde{Y} - \tilde{X} \hat{\Theta}\|_F^2 + \lambda_1 \|\hat{\Theta}\|_{\ell_1/\ell_2} \leq \frac{1}{n} \|\tilde{Y} - \tilde{X} \Theta_0\|_F^2 + \lambda_1 \|\Theta_0\|_{\ell_1/\ell_2}.
\]

Let \((A, B) = \text{tr}(A^TB)\) for any commensurate matrices \(A\) and \(B\). By writing \(\Delta := \hat{\Theta} - \Theta_0\) and noting that \(\tilde{Y} = \tilde{X} \Theta_0 + \tilde{X} L_0 + \tilde{\epsilon}\) with \(\tilde{\epsilon} = Q_{\lambda_2}^{1/2} \epsilon\), standard arguments yield

\[
\frac{1}{n} \|\tilde{X} \Delta\|_F^2 \leq \frac{2}{n} \|\tilde{X} L_0 + \tilde{\epsilon}, \tilde{X} \Delta\|_F + \lambda_1 \left( \|\Theta_0\|_{\ell_1/\ell_2} - \|\hat{\Theta}\|_{\ell_1/\ell_2} \right)
\]

\[
\leq \frac{2}{n} \|\tilde{X} L_0\|_F \|\tilde{X} \Delta\|_F + \frac{2}{n} \|\tilde{\epsilon}, \tilde{X} \Delta\|_F + \lambda_1 \left( \|\Theta_0\|_{\ell_1/\ell_2} - \|\Delta + \Theta_0\|_{\ell_1/\ell_2} \right)
\]

\[
\leq \frac{2}{n} \|\tilde{X} L_0\|_F \|\tilde{X} \Delta\|_F + \frac{2}{n} \max_{1 \leq j \leq p} \|\tilde{X}^T \tilde{\epsilon}\|_2 \|\Delta\|_{\ell_1/\ell_2} + \lambda_1 \left( \|\Theta_0\|_{\ell_1/\ell_2} - \|\Delta + \Theta_0\|_{\ell_1/\ell_2} \right) \quad \text{(B.6)}
\]
where we use Cauchy-Schwarz in the second inequality and the following display to derive the third inequality,

$$\langle \tilde{\varepsilon}, \tilde{X}\Delta \rangle = \left| \sum_{j=1}^{p} \tilde{X}_{j}\tilde{\varepsilon}\Delta_{j} \right| \leq \max_{1 \leq j \leq p} \| \tilde{X}_{j}\tilde{\varepsilon} \| \| \Delta \|_{\ell_{1}/\ell_{2}}.$$ 

On the event

$$\mathcal{E} := \left\{ \max_{1 \leq j \leq p} \| X_{j}^{T}Q_{\lambda_{2}}\varepsilon \| \leq \frac{n}{4} \cdot \lambda_{1} \right\},$$

by \( \| \tilde{X}_{j}\tilde{\varepsilon} \| = \| X_{j}^{T}Q_{\lambda_{2}}\varepsilon \| \), we further have

$$\frac{1}{n} \| \tilde{X}\Delta \|_{F}^{2} \left( 1 - \frac{2\| \tilde{X}L_{0}\|_{F}}{\| \tilde{X}\Delta \|_{F}} \right) \leq \lambda_{1} \left( \| \Theta_{0} \|_{\ell_{1}/\ell_{2}} - \| \Delta \|_{\ell_{1}/\ell_{2}} + \frac{1}{2} \| \Delta \|_{\ell_{1}/\ell_{2}} \right)$$

(B.8)

Notice that

$$\| \tilde{X}L_{0}\|_{F}^{2} = \text{tr} \left[ L_{0}^{T}X^{T}Q_{\lambda_{2}}XL_{0} \right] = n\lambda_{2}\text{tr} \left[ L_{0}^{T}\Sigma(\Sigma + \lambda_{2}I_{p})^{-1}L_{0} \right]$$

(B.9)

by using Fact 1. When \( \| \tilde{X}\Delta \|_{F} \leq 4\| \tilde{X}L_{0}\|_{F} \), we obtain the desired result from (B.9). It suffices to consider the case \( \| \tilde{X}\Delta \|_{F} \geq 4\| \tilde{X}L_{0}\|_{F} \). Display (B.8) then implies

$$\frac{1}{n} \| \tilde{X}\Delta \|_{F}^{2} \leq 2\lambda_{1} \left( \| \Theta_{0} \|_{\ell_{1}/\ell_{2}} - \| \Delta \|_{\ell_{1}/\ell_{2}} + \frac{1}{2} \| \Delta \|_{\ell_{1}/\ell_{2}} \right),$$

from which we conclude \( \Delta \in \mathcal{R}(1/2, \Theta_{0}, \lambda_{1}, \lambda_{2}) \) defined in (3.2). Invoking condition (3.1) with \( c = 1 \) gives

$$\frac{1}{\sqrt{n}} \| \tilde{X}\Delta \|_{F} \leq \frac{2\lambda_{1}}{\kappa_{1}(1/2, \Theta_{0}, \lambda_{1}, \lambda_{2})}.$$ 

Therefore, on the event \( \mathcal{E} \), by combining with (B.9), we have

$$\frac{1}{n} \| \tilde{X}\Delta \|_{F}^{2} \leq \max \left\{ 16\lambda_{2}\text{tr} \left[ L_{0}^{T}\Sigma(\Sigma + \lambda_{2}I_{p})^{-1}L_{0} \right], \frac{4\lambda_{2}^{2}}{\kappa_{1}^{2}(1/2, \Theta_{0}, \lambda_{1}, \lambda_{2})} \right\}.$$ 

(B.10)

Since the choice of \( \lambda_{1} \) in (3.4) together with Lemma 13 implies \( \mathbb{P}(\mathcal{E}) = 1 - \epsilon' \), we conclude the proof by invoking (3.4) in the above display.

\( \Box \)

**B.4 Proof of Theorem 5: convergence rate of \( \| \tilde{\Theta} - \Theta^{\ast} \|_{\ell_{1}/\ell_{2}} \)**

We work on the event \( \mathcal{E}_{B^{\ast}} := \{ \| \tilde{P} - P_{B^{\ast}} \|_{F} \leq c\xi_{n} \} \) for some constant \( c > 0 \). Define

$$\text{Rem}_{5} = C \left\{ \frac{1}{\sqrt{n}} \| XF^{\ast} \|_{op} + \Lambda_{1}^{1/2} \left( 1 + \frac{\sqrt{K}}{n} \right) \right\} \xi_{n}$$

(B.11)

for some constant \( C > 0 \). We will prove that for any \( \lambda_{s} \geq \lambda_{3} \), the solution \( \tilde{\Theta} \) from (2.8) satisfies

$$\frac{1}{\sqrt{n}} \| X\tilde{\Theta} - X\Theta^{\ast} \|_{F} \leq \max \left\{ 4\text{Rem}_{5}, \frac{3\lambda_{s}\sqrt{s}}{\kappa(s, s, 3)} \right\},$$

(B.12)

$$\| \tilde{\Theta} - \Theta^{\ast} \|_{\ell_{1}/\ell_{2}} \leq \max \left\{ \frac{\text{Rem}_{5}^{2}}{\lambda_{3}}, \frac{\lambda_{s}s_{s}}{\kappa^{2}(s, s, 4)} \right\}$$

(B.13)
with probability $1 - \varepsilon - 2e^{-c''K}$ for some constant $c'' > 0$. Then the result of Theorem 5 follows immediately by noting that $Rem_5 = \lambda_3 \sqrt{s_\ell} / \kappa(s_*, 4)$.

We proceed to prove (B.12) and (B.13). Pick any $\lambda_3 \geq \lambda_3$. Starting from (2.5), by writing $\Delta := \hat{\Theta} - \Theta^*$, standard arguments yield

$$\frac{1}{n} \|X\Delta\|_F^2 \leq \frac{2}{n} \langle Y\hat{P}^\perp - X\Theta^*, X\Delta \rangle + \lambda_3 \left(\|\Theta^*\|_{\ell_1/\ell_2} - \|\hat{\Theta}\|_{\ell_1/\ell_2}\right)$$

$$\leq \frac{2}{n} \langle E\hat{P}^\perp, X\Delta \rangle + \lambda_3 \left(\|\Delta S_r\|_{\ell_1/\ell_2} - \|\Delta S_c\|_{\ell_1/\ell_2}\right)$$

with $S_* := \{j \in [p] : \|\Theta_j\|_2 \neq 0\}$. Since

$$\langle E\hat{P}^\perp, X\Delta \rangle \leq \|\Delta\|_{\ell_1/\ell_2} \max_{1 \leq j \leq p} \|X_j E\hat{P}^\perp\|_2,$$

on the event $E'$ defined as

$$\left\{ \max_{1 \leq j \leq p} \|X_j E\hat{P}^\perp\|_2 \leq n\lambda_3 / 4 \right\} \cap \left\{ \frac{1}{\sqrt{n}} \left\| (XF^* + WB^*)\hat{P}^\perp - X\Theta^* \right\|_F \leq Rem_5 \right\},$$

by using $|\langle M, N \rangle| \leq \|M\|_F \|N\|_F$ for any commensurate matrices, we obtain

$$\frac{1}{n} \|X\Delta\|_F^2 \leq \frac{2}{\sqrt{n}} \|X\Delta\|_F \cdot Rem_5 + \frac{\lambda_3}{2} \left(3\|\Delta S_r\|_{\ell_1/\ell_2} - \|\Delta S_c\|_{\ell_1/\ell_2}\right). \quad (B.14)$$

By rearranging terms, we have

$$\frac{1}{n} \|X\Delta\|_F^2 \left(1 - \frac{2Rem_5}{\|X\Delta\|_F / \sqrt{n}}\right) \leq \frac{\lambda_3}{2} \left(3\|\Delta S_r\|_{\ell_1/\ell_2} - \|\Delta S_c\|_{\ell_1/\ell_2}\right).$$

When $\|X\Delta\|_F / \sqrt{n} \leq 4Rem_5$, (B.12) holds. When $\|X\Delta\|_F / \sqrt{n} \geq 4Rem_5$, we have

$$\frac{1}{n} \|X\Delta\|_F^2 \leq \lambda_3 \left(3\|\Delta S_r\|_{\ell_1/\ell_2} - \|\Delta S_c\|_{\ell_1/\ell_2}\right).$$

Hence $\Delta \in C(S_*, 3) \subseteq C(S_*, 4)$. Invoking (3.5) with $s = s_\ell$ and $\alpha = 3$ yields

$$\frac{1}{n} \|X\Delta\|_F^2 \leq 3\lambda_3 \|\Delta S_r\|_{\ell_1/\ell_2} \leq 3\lambda_3 \sqrt{s_\ell} \|\Delta S_r\|_F \leq \frac{3\lambda_3 \sqrt{s_\ell}}{\kappa(s_*, 3)} \left\{ \frac{1}{\sqrt{n}} \|X\Delta\|_F \right\},$$

which implies the first result on the event $E' \cap E_{B^*}$.

To show (B.13), note that $\kappa(s_*, 4) > 0$ and consider two cases:

1. When $\Delta \in C(S_*, 4)$, from the definition of $\kappa(s_*, 4)$, one has

$$\|\Delta\|_{\ell_1/\ell_2} = \|\Delta S_r\|_{\ell_1/\ell_2} + \|\Delta S_c\|_{\ell_1/\ell_2}$$

$$\leq 5\|\Delta S_r\|_{\ell_1/\ell_2}$$

$$\leq 5\sqrt{s_\ell} \|\Delta S_r\|_F$$

$$\leq \frac{5\sqrt{s_\ell}}{\kappa(s_*, 4)} \left\{ \frac{1}{\sqrt{n}} \|X\Delta\|_F \right\}. $$

49
(2) When $\Delta \notin C(S_*, 4)$, we have $4\|\Delta S_*\|_{\ell_1/\ell_2} < \|\Delta S_c^*\|_{\ell_1/\ell_2}$ by definition. Plugging this into display (B.14), we have

$$\frac{1}{n}\|X\Delta\|^2_F \leq \frac{2}{\sqrt{n}}\|X\Delta\|_F \cdot Rem_5 - \frac{\lambda_3}{8}\|\Delta S_c^*\|_{\ell_1/\ell_2}.$$ 

It implies $\|X\Delta\|_F/\sqrt{n} \leq 2Rem_5$ and

$$\lambda_3\|\Delta S_c^*\|_{\ell_1/\ell_2} \leq \frac{16}{\sqrt{n}}\|X\Delta\|_F \cdot Rem_5.$$ 

We thus obtain

$$\|\Delta\|_{\ell_1/\ell_2} \leq \frac{5}{4}\|\Delta S_c^*\|_{\ell_1/\ell_2} \leq \frac{20Rem_5}{\lambda_3} \frac{1}{\sqrt{n}}\|X\Delta\|_F \leq 40\frac{Rem_5^2}{\lambda_3}.$$ 

Combining these two cases and invoking (B.12) give the desired result on the event $E' \cap E_{B^*}$. Finally, invoking Lemmas 6 and 14 yield $P(E') \geq 1 - \epsilon - 2m^{-c^2K}$. This completes the proof. 

The following lemma gives that the probability of $E'$ on the event $E_{B^*}$. Recall that $Rem_5$ is defined in (B.11).

**Lemma 6.** Under conditions of Theorem 5, on the event $E_{B^*} = \{\|\hat{P} - P_{B^*}\|_F \leq c\xi_n\}$ for some constant $c > 0$, the following holds with probability $1 - 2e^{-cK}$ for some constant $c' > 0$.

$$\frac{1}{\sqrt{n}}\left\|\left(XF^* + WB^*\right)\hat{P} - X\Theta^*\right\|_F \leq Rem_5.$$ 

**Proof of Lemma 6.** By noting that $(XF^* + WB^*)P_{B^*} = X\Theta^*$, we have

$$\frac{1}{\sqrt{n}}\left\|\left(XF^* + WB^*\right)\hat{P} - X\Theta^*\right\|_F \leq \frac{1}{\sqrt{n}}\left\|\left(XF^* + WB^*\right)(\hat{P} - P_{B^*})\right\|_F \leq \frac{1}{\sqrt{n}}\left\|\left(XF^* + WB^*\right)\right\|_{op}\left\|\hat{P} - P_{B^*}\right\|_F \leq \frac{1}{\sqrt{n}}\left(\|XF^*\|_{op} + \|WB^*\|_{op}\right)c\xi_n,$$

where we have used $\|MN\|_F \leq \|M\|_{op}\|N\|_F$ for any commensurate matrices in the third line and the triangle inequality in the last line. For $WB^*$, invoking Lemma 15 gives

$$\frac{1}{n}\|WB^*\|_{op}^2 \leq \|(B^*)^T\Sigma WB^*\|_{op}\left[1 + C_{\gamma_w}\left(\sqrt{\frac{K}{n}} + \sqrt{\frac{K}{n}}\right)\right]$$

with probability $1 - 2e^{-cK}$ for some constants $c' > 0$ and $C_{\gamma_w}$ depending on $\gamma_w$ only. This completes the proof. 

}$\Box$
B.5 Proof of Theorem 6: convergence rate of $\|\hat{P}_{B^*} - P_{B^*}\|_F$ for homoscedastic case

Under Assumption 1, $L^*$ and $\Theta^*$ are identifiable and so is the row-sparsity $s_*$ of $\Theta^*$. Recall that $\hat{P}_{B^*} = \hat{U}^T \hat{U}$ and $P_{B^*} = UU^T$. Applying Theorem 2 in Yu et al. (2014) to $\Sigma_e$ and $\hat{\Sigma}_e$ with $d = s = K$ and $r = 1$ yields

$$
\|\hat{U}Q - U\|_F \leq \frac{2^{3/2}\|\hat{\Sigma}_e - \Sigma_e\|_F}{\lambda_K(\Sigma_e) - \lambda_{K+1}(\Sigma_e)} = \frac{2^{3/2}\|\hat{\Sigma}_e - \Sigma_e\|_F}{\lambda_K((B^*)^T \Sigma_W B^*)}
$$

(B.15)

for some orthogonal matrix $Q$. We also use the fact that $\Sigma_E = \tau^2 I_m$ and $\lambda_{K+1}(B^*) = 0$. Note that, for this $Q$,

$$
\|\hat{U} \hat{U}^T - UU^T\|_F \leq \|\hat{U}Q - U\|_F\|Q\|_2 \leq 2\|\hat{U}Q - U\|_F = 2\|\hat{U}Q - U\|_F.
$$

It then suffices to upper bound $\|\hat{\Sigma}_e - \Sigma_e\|_F$. Notice that

$$
\hat{\Sigma}_e - \frac{1}{n} \epsilon^T \epsilon = \frac{1}{n}(Y - X \hat{F})^T(Y - X \hat{F}) - \frac{1}{n} \epsilon^T \epsilon
$$

$$
= \frac{1}{n}(F - F^*)^T X^T (\hat{F} - F^*) + \frac{1}{n} \epsilon^T \epsilon X (F^* - \hat{F}).
$$

(B.16)

Recalling that (B.3), we have

$$
X \hat{F} - X F^* = P_{\lambda_2}(Y - X \Theta^*) - XL^* + Q_{\lambda_2}X(\hat{\Theta} - \Theta^*)
$$

$$
= P_{\lambda_2} \epsilon - Q_{\lambda_2}X L^* + Q_{\lambda_2}X(\hat{\Theta} - \Theta^*).
$$

By using formula $\|H^TH\|_F \leq \|H\|_F^2$ for any matrix $H$ and triangle inequality, we thus have

$$
\|\hat{\Sigma}_e - \Sigma_e\|_F \leq \frac{1}{n} \|X \hat{F} - X F^*\|_F^2 + \frac{2}{n} \|\epsilon^T X(\hat{F} - F^*)\|_F + \frac{1}{n} \|\epsilon^T \epsilon - \Sigma_e\|_F
$$

$$
\leq \frac{1}{n} \|X \hat{F} - X F^*\|_F^2 + \frac{2}{n} \|\epsilon^T P_{\lambda_2} \epsilon\|_F + \frac{2}{n} \|\epsilon^T Q_{\lambda_2} X L^*\|_F
$$

+ \frac{2}{n} \|\epsilon^T Q_{\lambda_2} X(\hat{\Theta} - \Theta^*)\|_F + \frac{1}{n} \|\epsilon^T \epsilon - \Sigma_e\|_F.
$$

We then study each terms on the right hand side. From Lemma 12, we have

$$
\frac{1}{n} \|\epsilon^T P_{\lambda_2} \epsilon\|_F \leq \frac{1}{n} \|P_\lambda^{1/2} \epsilon\|_F^2 \leq V_\epsilon \left(\sqrt{\text{tr}(P_{\lambda_2})} + \sqrt{2\|P_{\lambda_2}\|_{op} \log(m/\epsilon)}\right)^2
$$

$$
\frac{1}{n} \|\epsilon^T Q_{\lambda_2} X L^*\|_F \leq \sqrt{\frac{V_\epsilon \log(m/\epsilon)}{n}} \sqrt{\|Q_{\lambda_2}\|_{op} \cdot \text{Rem}_2(L^*)}
$$

with probability $1 - 2\epsilon$. To bound the fourth term, first notice that

$$
\frac{1}{n} \|\epsilon^T Q_{\lambda_2} X(\hat{\Theta} - \Theta^*)\|_F \leq \max_{1 \leq j \leq p} \frac{1}{n} \|X_j^T Q_{\lambda_2} \epsilon\|_2 \cdot \|\hat{\Theta} - \Theta^*\|_{\ell_1/\ell_2}.
$$
Indeed, by writing $\Delta = \hat{\Theta} - \Theta^*$, one has
\[
\|e^T Q_{\lambda_2} X \Delta\|^2_F = \sum_{\ell=1}^m \Delta_{i,\ell}^T X^T Q_{\lambda_2} e e^T Q_{\lambda_2} X \Delta_{\ell}
\leq \sum_{\ell=1}^m \sum_{i=1}^p |\Delta_{i,\ell}| \sum_{j=1}^p |\Delta_{j,\ell}| \max_{i,j} |X_i^T Q_{\lambda_2} e e^T Q_{\lambda_2} X_j|
\leq \sum_{i=1}^p \sum_{j=1}^p \|\Delta_{i,\ell}\|_2 \|\Delta_{j,\ell}\|_2 \max_{1 \leq j \leq p} \|X_j^T Q_{\lambda_2} e\|^2_2
= \|\Delta\|^2_{\ell_1/\ell_2} \max_{1 \leq j \leq p} \|X_j^T Q_{\lambda_2} e\|^2_2.
\]
Note that, on the event $\mathcal{E}$ defined in (B.7),
\[
|\hat{\Theta} - \Theta^*|_{\ell_1/\ell_2} \lesssim \frac{\lambda_1 s_*}{\kappa^2(s_*, 4)} + \frac{\text{Rem}_2(L^*)}{\lambda_1}
\]
from (B.23). Invoking (B.7) yields
\[
\frac{1}{n} \|e^T Q_{\lambda_2} X (\hat{\Theta} - \Theta^*)\|_F \lesssim \text{Rem}_2(L^*) + \frac{\lambda_1^2 s_*}{\kappa^2(s_*, 4)}
\]
with probability $1 - c'$. Finally, the last term can be upper bounded by invoking Lemma 16 as
\[
\mathbb{P}\left\{\left\|\frac{1}{n} e^T \vartheta - \Sigma_{\vartheta}\right\|_F \leq c V_{\vartheta} \left(\sqrt{\frac{\log m}{n}} \lor \sqrt{\frac{m}{n}}\right)\right\} \geq 1 - 2m^{-c'} \quad (B.17)
\]
for some constant $c, c' > 0$. Collecting terms and invoking Theorem 7 for $\|X \hat{F} - X F^*\|^2_F/n$ with $L_0 = L^*$ and $\Theta_0 = \Theta^*$ yield, after using $\|P_{\lambda_2}\|_{op} \leq 1$, $\|Q_{\lambda_2}\|_{op} \leq 1$ and Lemma 8 to simplify the results,
\[
\|\hat{\Sigma} - \Sigma\|_F \lesssim \text{Rem}_2(L^*) + \frac{\lambda_1^2 s_*}{\kappa^2(s_*, 4)} + \frac{V_{\vartheta} \log(m/\epsilon)}{n} \sqrt{\text{Rem}_2(L^*)} + V_{\vartheta} \left(\sqrt{\frac{\log m}{n}} \lor \sqrt{\frac{m}{n}}\right)
\leq \text{Rem}_2(L^*) + \frac{\lambda_1^2 s_*}{\kappa^2(s_*, 4)} + \frac{\text{tr}(P_{\lambda_2}) V_{\vartheta}}{n} + V_{\vartheta} \left(\sqrt{\frac{\log(m/\epsilon)}{n}} \lor \sqrt{\frac{m}{n}}\right) \quad (B.18)
\]
with probability $1 - 3\epsilon - \epsilon' - 2m^{-\epsilon'}$. Recall the eigen-decomposition of $\hat{\Sigma} = U \text{diag}(\sigma_1, \ldots, \sigma_p) U^T$ with $U = (u_1, \ldots, u_K)$. We have
\[
\text{tr}(P_{\lambda_2}) = \text{tr} \left[\frac{1}{n} X (\hat{\Sigma} + \lambda_2 I_p)^{-1} X^T\right] = \sum_{k=1}^q \frac{\sigma_k}{\sigma_k + \lambda_2},
\]
\[
\text{Rem}_2(L^*) = \sum_{k=1}^q \frac{\lambda_2 \sigma_k}{\sigma_k + \lambda_2} u_k^T (L^*)^T u_k \leq \frac{\lambda_2^3 \sigma_1}{\sigma_1 + \lambda_2} \|L^*\|^2_F. \quad (B.19)
\]
By invoking Lemma 8 with $s_0 = s_*$ together with the choice of $\lambda_1$ as (3.4), we further have
\[
\frac{\lambda_1^2 s_*}{\kappa^2(s_*, 4)} \lesssim \frac{\sigma_1 + \lambda_2}{\lambda_2^2} \frac{\lambda_1^2 s_*}{\kappa^2(s_*, 4)} \lesssim \frac{\lambda_2 (\sigma_1 + \lambda_2)}{\sigma_q + \lambda_2} \|s_* V_{\vartheta} \|_2 \log N \quad (B.20)
\]
Take $\epsilon = m^{-\epsilon'}$ and use $\log m \leq n$ in (B.18) to complete the proof.
B.6 Lemma 7 used in the proof of Theorem 6

The following lemma provides the rate of $||\hat{\Theta} - \Theta^*||_{\ell_1/\ell_2}$ where $\hat{\Theta}$ is obtained in (2.3). Furthermore, its proof reveals that Lemma 7 holds by replacing $\lambda_1$ by any $\tilde{\lambda}_1 \geq \lambda_1$.

**Lemma 7.** Under conditions of Corollary 4, choose $\lambda_1$ as (3.4) and any $\lambda_2 \geq 0$ such that $P_{\lambda_2}$ exists. Assume $\tilde{\kappa}(s_*, 4) > 0$ holds. With probability $1 - \epsilon - \epsilon'$,

$$||\hat{\Theta} - \Theta^*||_{\ell_1/\ell_2} \lesssim \frac{\lambda_1 s_*}{\tilde{\kappa}^2(s_*, 4)} + \frac{\text{Rem}_2(L^*)}{\lambda_1}$$

(B.21)

where $\text{Rem}_2(L^*)$ is defined in Theorem 3.

**Proof.** We prove (B.21) by working on the event $\mathcal{E}$ defined in (B.7). Note that $\Theta^*$ and $L^*$ are identifiable. From (B.10) in the proof of Lemma 5, by taking $L_0 = L^*$ and $\Theta_0 = \Theta^*$,

$$\frac{1}{n} ||\tilde{X}(\hat{\Theta} - \Theta^*)||_F^2 \leq \max \left\{ 16 \lambda_2 \text{tr} \left[ (L^*)^T \Sigma (\Sigma + \lambda_2 I_p)^{-1} L^* \right], \frac{4 \lambda_1^2}{\kappa_1^2(1/2, \Theta^*, \lambda_1, \lambda_2)} \right\}$$

$$\lesssim \text{Rem}_2(L^*) + \frac{\lambda_1^2 s_*}{\kappa_2^2(s_*, 3)},$$

(B.22)

by invoking the first result of Lemma 8 with $s_0 = s_*$ and $c = 1/2$. Write $\Delta := \hat{\Theta} - \Theta^*$ and consider two cases.

1. When $\Delta \in \mathcal{C}(S_*, 4)$ with $S_* := \{ j \in [p] : ||\Theta_j||_2 \neq 0 \}$, it follows from the definitions of $\mathcal{C}(S_*, 4)$ and $\tilde{\kappa}(s_*, 4)$ that

$$||\Delta||_{\ell_1/\ell_2} \leq 5 ||\Delta_{S_*}||_{\ell_1/\ell_2} \leq 5 \sqrt{s_*} ||\Delta_{S_*}||_F \leq 5 \sqrt{s_*} \frac{1}{\tilde{\kappa}_1(s_*, 4)} \frac{1}{\sqrt{n}} ||\tilde{X}\Delta||_F.$$

2. When $\Delta \notin \mathcal{C}(S_*, 4)$, it implies $||\Delta_{S^c_*}||_{\ell_1/\ell_2} > 4 ||\Delta_{S_*}||_{\ell_1/\ell_2}$. From (B.6), by invoking $\mathcal{E}$, we obtain

$$\frac{1}{n} ||\tilde{X} \Delta||_F^2 \leq \frac{2}{n} ||\tilde{X} L^*||_F ||\tilde{X} \Delta||_F^2 + \frac{\lambda_1}{2} (3 ||\Delta_{S_*}||_{\ell_1/\ell_2} - ||\Delta_{S^c_*}||_{\ell_1/\ell_2})$$

$$\leq \frac{2}{n} ||\tilde{X} L^*||_F ||\tilde{X} \Delta||_F^2 - \frac{\lambda_1}{8} ||\Delta_{S^c_*}||_{\ell_1/\ell_2}.$$

This implies $||\tilde{X} \Delta||_F \leq 2 ||\tilde{X} L^*||_F$ and

$$||\Delta_{S^c_*}||_{\ell_1/\ell_2} \leq \frac{16}{\lambda_1} \frac{1}{n} ||\tilde{X} L^*||_F ||\tilde{X} \Delta||_F \leq \frac{32}{\lambda_1} \frac{1}{n} ||\tilde{X} L^*||_F^2.$$

Using $||\Delta_{S^c_*}||_{\ell_1/\ell_2} > 4 ||\Delta_{S_*}||_{\ell_1/\ell_2}$ again yields

$$||\Delta||_{\ell_1/\ell_2} \leq \frac{5}{4} ||\Delta_{S^c_*}||_{\ell_1/\ell_2} \leq \frac{40}{\lambda_1} \frac{1}{n} ||\tilde{X} L^*||_F^2.$$

Combining these two cases and invoking (B.9) with $L_0 = L^*$ and (B.22) give

$$||\Delta||_{\ell_1/\ell_2} \lesssim \frac{\sqrt{s_*}}{\tilde{\kappa}(s_*, 4)} \left( \sqrt{\text{Rem}_2(L^*)} + \frac{\lambda_1 \sqrt{s_*}}{\tilde{\kappa}(s_*, 3)} \right) + \frac{\text{Rem}_2(L^*)}{\lambda_1}$$

$$\lesssim \frac{\lambda_1 s_*}{\tilde{\kappa}^2(s_*, 4)} + \frac{\text{Rem}_2(L^*)}{\lambda_1}$$

(B.23)

where we also use $\tilde{\kappa}(s_*, 3) \geq \tilde{\kappa}(s_*, 4)$ to derive the last inequality. We then conclude the proof by recalling that $\mathbb{P}(\mathcal{E}) = 1 - \epsilon'$. \hfill \Box
B.7 Proof of Theorem 9: convergence rate of $\|\tilde{P}_{B^*} - P_{B^*}\|_F$ for heteroscedastic case

Let $M = (B^*)^T \Sigma W B^*$. From (B.16) and by using $\Sigma_e = M + \Sigma_E$, one has

$$\hat{\Sigma}_e = M + \Delta_1 + \Delta_2$$

where $\Delta_1 = \Sigma_E = \text{diag}(\tau^2_1, \ldots, \tau^2_m)$ and

$$\Delta_2 = \frac{1}{n}(\hat{F} - F^*)^T X^T X(\hat{F} - F^*) + \frac{1}{n}(F^* - \hat{F})^T X^T \varepsilon + \frac{1}{n} \varepsilon^T X(F^* - \hat{F}) + \frac{1}{n} \varepsilon^T \varepsilon - \Sigma_e.$$

We aim to apply Theorem 12 with $N = \hat{\Sigma}_e$, $M = M$, $Z = \Delta_1 + \Delta_2$, $\hat{V} = \hat{U}$ and $V = U$. Observe that $\|\Gamma(\Delta_1)\|_F = 0$, that is, the off-diagonal elements of $\Delta_1$ are zero, and $\|\Delta_2\|_F = \|\hat{\Sigma}_e - \Sigma_e\|_F$ from (B.16). Invoking (B.18) yields

$$\mathbb{P}\{\|\Delta_2\|_F \leq \text{Rem}(P_{B^*}) \cdot \lambda_K(M)\} \geq 1 - \epsilon' - 5m^{-\epsilon'}$$

(B.24)

with $\text{Rem}(P_{B^*})$ defined in (3.14). We then work on the event that the above display holds. Recall that $\Gamma(\Delta_2)$ denotes the matrix with off-diagonal elements equal to $\Delta_2$ and diagonal elements equal to zero. Since $\text{Rem}(P_{B^*}) \leq c\sqrt{K}$ implies $\|\Gamma(\Delta_2)\|_F \leq \|\Delta_2\|_F \leq c\sqrt{K} \lambda_K(M)$, in conjunction with condition (4.2), an application of Theorem 12 with $N = \hat{\Sigma}_e$, $M = M$, $Z = \Delta_1 + \Delta_2$, $\hat{V} = \hat{U}$ and $V = U$ gives

$$\|\sin \Theta(\hat{U}, U)\|_F \leq \frac{\|\Gamma(\Delta_2)\|_F}{\lambda_K(M)} \wedge \sqrt{K} \leq \frac{\|\Delta_2\|_F}{\lambda_K(M)}.$$

Finally, using the inequality

$$\|\tilde{P}_{B^*} - P_{B^*}\|_F = \|\hat{U} \hat{U}^T - UU^T\|_F \leq 2\|\sin \Theta(\hat{U}, U)\|_F.$$

and (B.24) again concludes the proof.

B.8 Proof of Theorem 10: consistency of using PCA to estimate $P_{B^*}$ in the presence of heteroscedasticity

The proof follows the same arguments as that of Theorem 6. The first difference is to apply Theorem 2 in Yu et al. (2014) to $\hat{\Sigma}_e$ and $\Pi := B^T \Sigma W B^* + \tilde{\tau}^2 I_m$ with $d = s = K$ and $r = 1$ to obtain

$$\|\hat{U} Q - U\|_F \leq \frac{2^{3/2}\|\hat{\Sigma}_e - \Pi\|_F}{\lambda_K(\Pi) - \lambda_{K+1}(\Pi)} = \frac{2^{3/2}\|\hat{\Sigma}_e - \Pi\|_F}{\lambda_K(\Pi)}.$$

The second difference from the proof of Theorem 6 is to upper bound the numerator as

$$\|\hat{\Sigma}_e - \Pi\|_F \leq \|\hat{\Sigma}_e - \Sigma_e\|_F + \|\Sigma_e - \Pi\|_F = \|\hat{\Sigma}_e - \Sigma_e\|_F + \|\Sigma_E - \tilde{\tau}^2 I_m\|_F$$

by adding and subtracting $\Sigma_e$ and using $\Sigma_e = \Pi + \Sigma_E - \tilde{\tau}^2 I_m$. Since the results in Theorem 3 still hold in the heteroscedastic case, $\|\hat{\Sigma}_e - \Sigma_e\|_F$ can be bounded by (B.18). Then the proof is completed by using

$$\|\Sigma_E - \tilde{\tau}^2 I_m\|_F^2 = \sum_{j=1}^{m} (\tau_j^2 - \tilde{\tau}^2)^2.$$
B.9 Proof of Theorem 11: selection of $K$

Recall that $M = (B^*)^T \Sigma W B^*$ and its eigenvalues are $\Lambda_1 \geq \Lambda_2 \geq \cdots \geq \Lambda_K$ and $\Lambda_j = 0$ for $j > K$. Recall that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_m$ are the eigenvalues of $\hat{\Sigma}_e$. By Weyl's inequality, we have

$$|\hat{\lambda}_j - \Lambda_j| \leq \|\hat{\Sigma}_e - M\|_{\text{op}}$$

for all $1 \leq j \leq m$. We work on the event $\{\|\hat{\Sigma}_e - \Sigma_e\|_F \lesssim \text{Rem}(P_{B^*})\Lambda_K\}$ which, from the proof of Theorem 6, holds with probability $1 - \epsilon' - 5m^{-c}$ for some constant $c > 0$. Note that

$$\|\hat{\Sigma}_e - M\|_{\text{op}} \leq \|\hat{\Sigma}_e - \Sigma_e\|_F + \|\Sigma_e\|_{\text{op}} \lesssim \text{Rem}(P_{B^*})\Lambda_K + \max_{1 \leq j \leq m} \tau_j^2. \tag{B.25}$$

We thus conclude

$$|\hat{\lambda}_j - \Lambda_j| \lesssim \max_{1 \leq j \leq m} \tau_j^2 + \text{Rem}(P_{B^*})\Lambda_K$$

for $1 \leq j \leq m$. Since (b) of Assumption 3 implies $\Lambda_j \approx m$ for $1 \leq j \leq K$, by also using $\text{Rem}(P_{B^*}) = o(1)$ and $\max_j \tau_j^2 = O(1)$, we have $\hat{\lambda}_j \approx m$. This concludes $\hat{\lambda}_{j+1}/\hat{\lambda}_j \approx 1$ for $1 \leq k \leq K - 1$. On the other hand, since $\hat{\lambda}_{K+1} = O(\max_j \tau_j^2 + \text{Rem}(P_{B^*})\Lambda_K)$, we further obtain $\hat{\lambda}_{K+1}/\hat{\lambda}_K = O(\max_j \tau_j^2/m + \text{Rem}(P_{B^*}))$. This completes the proof. \square

C Auxiliary proofs and technical lemmas

C.1 Lemmas used in Remark 1

We establish the connection between the impact factor defined in (3.1) and the RE conditions of $\tilde{X}$ and $X$. Recall that $M = n^{-1}X^T Q_{\lambda_2}^2 X$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q > 0$ are the non-zero eigenvalues of $\hat{\Sigma}$ with $q := \text{rank}(X)$.

Lemma 8. For any given $\Theta_0$ with row-sparsity $s_0$ and any constant $c \in (0, 1)$, one has

$$\kappa_1(c, \Theta_0, \lambda_1, \lambda_2) \geq \frac{\bar{\kappa}(s_0, \alpha_c)}{(1 + c)\sqrt{s_0}} \geq \sqrt{\frac{\lambda_2}{\sigma_1 + \lambda_2}} \cdot \frac{\kappa(s_0, \alpha_c)}{(1 + c)\sqrt{s_0}}$$

with $\alpha_c = (1 + c)/(1 - c)$, and

$$\max_{1 \leq j \leq p} M_{jj} \leq \max_{1 \leq j \leq p} \hat{\Sigma}_{jj} \left( \frac{\lambda_2}{\sigma_1 + \lambda_2} \right)^2.$$

As a result, we have

$$\frac{\max_{1 \leq j \leq p} M_{jj}}{\kappa_1^2(1/2, \Theta_0, \lambda_1, \lambda_2)} \leq \frac{9s_0}{4\kappa^2(s_0, 3)} \max_j \hat{\Sigma}_{jj} \cdot \frac{\lambda_2(\sigma_1 + \lambda_2)}{(\sigma_1 + \lambda_2)^2}. $$

Proof. We first prove

$$\kappa_1(c, \Theta_0, \lambda_1, \lambda_2) \geq \frac{\bar{\kappa}(s_0, \alpha_c)}{(1 + c)\sqrt{s_0}}.$$
Observe that $\mathcal{R}(c, \Theta_0, \lambda_1, \lambda_2) \subseteq \mathcal{C}(S, \alpha_c)$ for any $|S| \leq s_0$, $c \in (0, 1)$ and $\alpha = (1 + c)/(1 - c)$. Indeed, for any $\Delta \in \mathcal{R}(c, \Theta_0, \lambda_1, \lambda_2)$, $|S| \leq s_0$ and $c \in (0, 1)$,

$$
0 \leq \|\Theta_0\|_{\ell_1/\ell_2} - \|\Theta_0 + \Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2}
\leq \|\Delta S\|_{\ell_1/\ell_2} - \|\Delta S\|_{\ell_1/\ell_2} + c\|\Delta S\|_{\ell_1/\ell_2} + c\|\Delta S\|_{\ell_1/\ell_2}
= (1 + c)\|\Delta S\|_{\ell_1/\ell_2} - (1 - c)\|\Delta S\|_{\ell_1/\ell_2},
$$

which implies $\Delta \in \mathcal{C}(S, \alpha_c)$. Note the above display also implies

$$
\|\Theta_0\|_{\ell_1/\ell_2} - \|\Theta_0 + \Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2} \leq \sqrt{s_0}(1 + c)\|\Delta S\|_F.
$$

We thus have

$$
\kappa(s_0, \alpha_c) = \min_{S \subseteq \{1, \ldots, p\}} \min_{\Delta \in \mathcal{C}(S, \alpha)} \frac{\|\tilde{X}\Delta\|_F/\sqrt{n}}{\|\Delta S\|_F}
\leq \min_{S \subseteq \{1, \ldots, p\}} \min_{\Delta \in \mathcal{C}(S, \alpha)} \frac{(1 + c)\sqrt{s_0} \cdot \|\tilde{X}\Delta\|_F/\sqrt{n}}{\|\Theta_0\|_{\ell_1/\ell_2} - \|\Theta_0 + \Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2}}
\leq \frac{(1 + c)\sqrt{s_0} \cdot \|\tilde{X}\Delta\|_F/\sqrt{n}}{\|\Theta_0\|_{\ell_1/\ell_2} - \|\Theta_0 + \Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2}}
= (1 + c)\sqrt{s_0}\kappa_1(c, \Theta_0, \lambda_1, \lambda_2).
$$

We then prove the second inequality of the first statement. Since

$$
\frac{1}{n}\|\tilde{X}\Delta\|_F^2 = \text{tr} \left[ \Delta^T \frac{1}{n} X^T Q_{\lambda_2} X \Delta \right]
= \lambda_2 \text{tr} \left[ \Delta^T \tilde{\Sigma}(\tilde{\Sigma} + \lambda_2 I_p)^{-1} \Delta \right]
\geq \frac{\lambda_2}{\|\tilde{\Sigma}\|_{op} + \lambda_2} \text{tr} \left( \Delta^T \tilde{\Sigma} \Delta \right)
$$

by using Fact 1 in the second line, it follows that

$$
\kappa^2(s, \alpha) \geq \frac{\lambda_2}{\sigma_1 + \lambda_2} \kappa^2(s, \alpha).
$$

We then show the second statement. From Fact 1 and $\tilde{\Sigma} = U \text{diag}(\sigma_1, \ldots, \sigma_q) U^T$, we have

$$
M = \lambda_2^2(\tilde{\Sigma} + \lambda_2 I_p)^{-1}\tilde{\Sigma}(\tilde{\Sigma} + \lambda_2 I_p)^{-1} = \lambda_2^2 UDU^T = \lambda_2^2 \tilde{\Sigma}^{1/2}(\tilde{\Sigma} + \lambda_2 I_p)^{-2} \tilde{\Sigma}^{1/2}
$$

with $D$ being diagonal and $D_{kk} = \sigma_k/(\sigma_k + \lambda_2)^2$ for $1 \leq k \leq q$. This implies

$$
\max_j M_{jj} \leq \max_j \tilde{\Sigma}_{jj} \left( \frac{\lambda_2}{\sigma_q + \lambda_2} \right)^2
$$

which, in conjunction with the previous result of $\kappa_1(c, \Theta_0, \lambda_1, \lambda_2)$ with $c = 1/2$, gives

$$
\frac{\max_{1 \leq i \leq p} M_{jj}}{\kappa_1^2(1/2, \Theta_0, \lambda_1, \lambda_2)} \leq \frac{g_{s_0}}{4\kappa^2(s_0, 3)} \max_j \tilde{\Sigma}_{jj} \cdot \frac{\lambda_2(\sigma_1 + \lambda_2)}{(\sigma_q + \lambda_2)^2}.
$$

This completes the proof. □
C.2 Proof of Corollary 4 and Remark 2

We first prove the following lemma from which Corollary 4 follows immediately. By taking \( q = p, \sigma_k = 1 \) for all \( 1 \leq k \leq q \) and \( \kappa(s_*, 3) = 1 \), the bound in (3.8) of Remark 2 follows from Corollary 4.

**Lemma 9.** Let \( \text{Rem}_1, \text{Rem}_2(L^*) \) and \( \text{Rem}_3(\Theta^*) \) be defined in Theorems 3. One has

\[
\text{Rem}_1 \lesssim \left\| \sum_{1 \leq k \leq q} \left( \frac{\sigma_k}{\sigma_k + \lambda_2} \right)^2 + \max_{1 \leq k \leq q} \left( \frac{\sigma_k}{\sigma_k + \lambda_2} \right)^2 \log(m/e) \right\| V_e / n,
\]

\[
\text{Rem}_2(L^*) \lesssim \frac{\lambda_2 \sigma_1}{\sigma_1 + \lambda_2} \|L^*\|^2_F,
\]

\[
\text{Rem}_3(\Theta^*) \lesssim \max_{1 \leq j \leq p} \left\| \frac{\lambda_2}{\sigma_q + \lambda_2} \left( \sigma_1 + \lambda_2 \right) \left( 1 + \frac{\log(p/e')}{r_e}(\Gamma_e) \right) \right\| \sum_{1 \leq j \leq p} \frac{s_* V_e}{\kappa^2(s_*, 4) \cdot n}
\]

and \( \|Q_{\lambda_2}\|_{op} \leq \lambda_2 / (\sigma_q + \lambda_2) \).

**Proof.** Recall that \( \hat{\Sigma} = U \text{diag}(\sigma_1, \ldots, \sigma_q) U^T \) with \( U = (u_1, \ldots, u_K) \) and \( P_{\lambda_2} = X(X^TX + n \lambda_2 I_p)^{-1}X^T \). The first result follows by observing

\[
\text{tr}(P_{\lambda_2}^2) = \text{tr} \left[ \frac{1}{n} X (\hat{\Sigma} + \lambda_2 I_p)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda_2 I_p)^{-1} X^T \right] = \sum_{k=1}^{q} \left( \frac{\sigma_k}{\sigma_k + \lambda_2} \right)^2,
\]

\[
\|P_{\lambda_2}\|_{op} = \left\| (\hat{\Sigma} + \lambda_2 I_p)^{-1/2} \frac{1}{n} X^T X (\hat{\Sigma} + \lambda_2 I_p)^{-1/2} \right\|_{op} \leq \frac{\sigma_1}{\sigma_1 + \lambda_2}.
\]

By noting that

\[
\text{tr} \left( (L^*)^T \hat{\Sigma} (\hat{\Sigma} + \lambda_2 I_p)^{-1} L^* \right) = \sum_{k=1}^{q} \frac{\sigma_k}{\sigma_k + \lambda_2} (L^*)^T u_k u_k^T L^* \leq \frac{\sigma_1}{\sigma_1 + \lambda_2} \|L^*\|^2_F,
\]

with \( U = (u_1, \ldots, u_q) \), the second result follows. The bound of \( \text{Rem}_3(\Theta^*) \) can be derived from Lemma 8 with \( s_0 = s_* \). Finally, since \( \|P_{\lambda_2}\|_{op} \geq \sigma_q / (\lambda_2 + \sigma_q) \), we immediately have

\[
\|Q_{\lambda_2}\|_{op} \leq 1 - \|P_{\lambda_2}\|_{op} \leq \frac{\lambda_2}{\sigma_q + \lambda_2}.
\]

The proof is complete. \( \square \)

C.3 Proof of Corollary 7, Remarks 4 and 5

**Proof of Corollary 7.** By inspecting the proof of Theorem 3, we can change the logarithmic factors in Theorem 5 and Theorem 6 to \( \log(N) \) with \( N = n \lor m \lor p \). The resulting probabilities will tend to 1 as \( n \rightarrow \infty \).

We first upper bound \( \text{Rem}(P_{B^*}; \lambda_2) := \text{Rem}(P_{B^*}) \) defined in (3.14). Here we write the dependency on \( \lambda_2 \) explicitly. Recall that

\[
\text{Rem}(P_{B^*}; \lambda_2) = \frac{1}{\lambda_K} \left\{ V_e \sqrt{\frac{\log m}{n}} + \frac{\lambda_2 \sigma_1}{\lambda_2 + \sigma_1} \|L^*\|^2_F + \sum_{k=1}^{q} \frac{\sigma_k}{\sigma_k + \lambda_2} \frac{V_e}{n} \right. \\
+ \left. \frac{\lambda_2}{\sigma_q + \lambda_2} \Sigma_{ij} \left( 1 + \frac{\log(p/e')}{r_e(\Gamma_e)} \right) \sum_{1 \leq j \leq p} \frac{s_* V_e}{\kappa^2(s_*, 4) \cdot n} \right\}.
\]
Further recalling that (3.3), we have \( V_\varepsilon \approx (K\Lambda_1\gamma_w^2 + m\gamma_e^2) \), under part (b) of Assumption 3. This implies
\[
\frac{V_\varepsilon}{K} \approx K\gamma_w^2 + \gamma_e^2 = O(K). \tag{C.3}
\]
From \( \kappa^{-1}(s, 4) = O(1) \) in part (a) of Assumption 3 and \( \hat{\Sigma}_{jj} = 1 \) for all \( 1 \leq j \leq p \), we conclude
\[
\text{Rem}(P_{B^*}; \lambda_2) \lesssim \frac{\lambda_2 \sigma_1}{\sigma_1 + \lambda_2} \frac{\|L^*\|_F^2}{m} + \frac{K^2}{n} \sum_{k=1}^{q} \frac{\sigma_k}{\lambda_k + \lambda_2} + \frac{\lambda_2(\sigma_1 + \lambda_2)}{(\sigma_1 + \lambda_2)^2} \cdot \frac{K_s}{n} + \frac{K}{\sqrt{n}}. \tag{C.4}
\]
We then prove
\[
\min_{\lambda_2} \text{Rem}(P_{B^*}; \lambda_2) \lesssim \min \left\{ \frac{\lambda_2 \sigma_1}{\sigma_1 + \lambda_2} \frac{\|L^*\|_F^2}{m} + \frac{K_s}{n}, \frac{qK}{n}, \sqrt{\frac{\|L^*\|_F^2}{m} \cdot \frac{(p + \sigma_1 s_*) K}{n}} + \frac{K_s}{n} \right\} + \frac{K}{\sqrt{n}}. \tag{C.5}
\]
To prove the first bound, by choosing \( \lambda_2 \to \infty \) in (C.4), we have
\[
\min_{\lambda_2} \text{Rem}(P_{B^*}) \lesssim \sigma_1 \frac{\|L^*\|_F^2}{m} + \frac{s_* K}{n} + \frac{K}{\sqrt{n}}.
\]
To prove the second bound, take \( \lambda_2 \to 0 \) in (C.4) to obtain
\[
\min_{\lambda_2} \text{Rem}(P_{B^*}) \lesssim \frac{qK}{n} + \frac{K}{\sqrt{n}}.
\]
Finally, from \( \sum_k \sigma_k = \text{tr}(\hat{\Sigma}) = p \), display (C.4) yields
\[
\text{Rem}(P_{B^*}; \lambda_2) \lesssim \lambda_2 \frac{\|L^*\|_F^2}{m} + \frac{pK}{\lambda_2 n} + \frac{\sigma_1 + \lambda_2 s_* K}{\lambda_2 n} + \frac{K}{\sqrt{n}}
\]
\[
= \lambda_2 \frac{\|L^*\|_F^2}{m} + \frac{(p + \sigma_1 s_*) K}{\lambda_2 n} + \frac{s_* K}{\lambda_2 n} + \frac{K}{\sqrt{n}}.
\]
Optimizing the above display over \( \lambda_2 \) yields
\[
\lambda_2^2 = \left( \frac{pK}{n} + \frac{\sigma_1 s_* K}{n} \right) \frac{m}{\|L^*\|_F^2}
\]
such that
\[
\min_{\lambda_2} \text{Rem}(P_{B^*}; \lambda_2) \lesssim \sqrt{\left( \frac{pK}{n} + \frac{\sigma_1 s_* K}{n} \right) \frac{\|L^*\|_F^2}{m} + \frac{s_* K}{n} + \frac{K}{\sqrt{n}}}
\]
We thus have proved (C.5).

We proceed to upper bound \( \|XF^*\|_{op}/\sqrt{n} \) by
\[
\frac{1}{\sqrt{n}} \|XL^*\|_{op} + \frac{1}{\sqrt{n}} \|X\Theta^*\|_{op} = O\left(\sqrt{m + \sqrt{m + s_*}}\right)
\]
where we use Assumption 3 in conjunction with
\[
\frac{1}{\sqrt{n}} \|X\Theta^*\|_{op} = \frac{1}{\sqrt{n}} \|Xs_{\Theta^*S_\Theta^*}\|_{op} \leq \left\|\hat{\Sigma}_{S_\Theta^*S_\Theta^*}\right\|_{op}^{1/2} \|\Theta^*\|_{op} = O\left(\sqrt{m + s_*}\right).
\]


58
by recalling that $S_\ast = \{ j \in [p] : \| \Theta_j^\ast \|_2 \neq 0 \}$. Using $K = O(n)$ concludes

$$
\frac{1}{\sqrt{n}} \| X F^\ast \|_{op} + \Lambda_1^{1/2} \left( 1 + \sqrt{\frac{K \log m}{n}} \right) \approx \sqrt{m + s_\ast}.
$$

(C.6)

On the other hand, we have

$$
\frac{\bar{\lambda}_3 s_\ast}{\kappa^2(s_\ast, 4)} \leq s_\ast \sqrt{\frac{m}{n}}.
$$

(C.7)

This completes the proof of Corollary 7. □

**Proof of Remark 4.** When $p = q < n$, $s_\ast \asymp p$ and $\sigma_1 = O(1)$, from Corollary 7, we have

$$
\text{Err}(P_{B^\ast}) = \min \left\{ \sigma_1 \frac{\| L^\ast \|_F^2}{m}, pK \frac{n}{n}, \sqrt{\frac{\| L^\ast \|_F^2}{m}}, \sqrt{\frac{(p + \sigma_1 s_\ast)K}{n}} + \frac{K}{\sqrt{n}} \right\} + \frac{K}{\sqrt{n}}
$$

$$
\leq \min \left\{ \frac{\| L^\ast \|_F^2}{m}, \frac{pK}{n}, \sqrt{\frac{\| L^\ast \|_F^2}{m}} + \frac{pK}{n} \right\} + \frac{K}{\sqrt{n}}
$$

Using $p \asymp s_\ast$ gives

$$
\sqrt{\frac{s_\ast (m + s_\ast)}{m}} \cdot \text{Err}(P_{B^\ast}) \approx \sqrt{\frac{p(m + p)}{m}} \left( \frac{pK}{n} + \frac{K}{\sqrt{n}} \right).
$$

The result of $p < n$ then follows by (C.7) and $K = O(\sqrt{p} \land \sqrt{n/p})$.

To prove the result of $p \geq n$, note that

$$
\text{Err}(P_{B^\ast}) \leq \min \left\{ \sigma_1 \frac{\| L^\ast \|_F^2}{m}, \sqrt{\frac{\| L^\ast \|_F^2}{m}}, \sqrt{\frac{(p + \sigma_1 s_\ast)K}{n}} + \frac{K}{\sqrt{n}} \right\} + \frac{K}{\sqrt{n}}.
$$

Condition $K = O(\sqrt{s_\ast} \land \sqrt{n/s_\ast} \land m)$ implies

$$
\sqrt{\frac{s_\ast (m + s_\ast)}{m}} \left( \frac{K}{\sqrt{n}} + \frac{K s_\ast}{n} \right) = \begin{cases} 
O \left( s_\ast / \sqrt{n} \right) & \text{if } s_\ast = O(m); \\
O \left( s_\ast / \sqrt{n} + (s_\ast / \sqrt{n})^2 \right) & \text{if } m = O(1).
\end{cases}
$$

The desired result then follows. □

**Proof of Remark 5.** From Lemma 7, we have

$$
\| \hat{\Theta} - \Theta^\ast \|_{\ell_1/\ell_2} = O_p \left( \frac{\tilde{\lambda}_1 s_\ast}{\kappa^2(s_\ast, 4)} + \frac{\text{Rem}_2(L^\ast)}{\tilde{\lambda}_1} \right)
$$

for any $\tilde{\lambda}_1 \geq \lambda_1$ where $\lambda_1$ defined in (3.4). For a suitable choice of $\tilde{\lambda}_1$, we can deduce that

$$
\| \hat{\Theta} - \Theta^\ast \|_{\ell_1/\ell_2} = O_p \left( \frac{\lambda_1 s_\ast}{\kappa^2(s_\ast, 4)} + \sqrt{\frac{s_\ast \text{Rem}_2(L^\ast)}{\kappa^2(s_\ast, 4)}} \right)
$$

By Lemma 8 and (B.19) together with $V_e = O(Km)$ and $[\kappa(s_\ast, 4)]^{-1} = O(1)$ under Assumption 3, we have

$$
\frac{1}{\sqrt{m}} \| \hat{\Theta} - \Theta^\ast \|_{\ell_1/\ell_2} \approx \frac{\sigma_1 + \lambda_2}{\sigma_2} \cdot s_\ast \sqrt{K} \frac{\sqrt{m}}{\sqrt{n}} + \sqrt{s_\ast} \frac{\| L^\ast \|_F^2}{m}.
$$

The result then follows by taking $\lambda_2 \geq \sigma_1$. □
C.4 Lemma used in Section 3.3

Lemma 10. Recall that $L^* = A^*B^*$ with $A^*$ defined in (1.3). Suppose $\Sigma_i^{-1/2}X_i$ are i.i.d. $\gamma_X$ sub-Gaussian random vectors for $1 \leq i \leq n$. Further assume $\|\Sigma_Z\|_{op} = O(1)$, $[\lambda_K(\Sigma_W)]^{-1} = O(1)$, $K = O(n)$ and part (b) of Assumption 3 holds. Then

$$\frac{1}{n}\|XL^*\|_{op}^2 = O_p(m).$$

Proof. Without loss of generality, assume $E[Z] = 0$ and $E[X] = 0$. Recall from (1.3) that

$$A^* = \Sigma^{-1}\Sigma_XZ$$

where we write $\Sigma_{XZ} = \text{Cov}(X, Z)$. Since $\Sigma_i^{-1/2}X_i$ are i.i.d. $\gamma_X$ sub-Gaussian, one has $(A^*)^T X_i$ is $\gamma_X\sqrt{\|\Sigma_{XZ}\Sigma_i^{-1}\Sigma_{XZ}\|_{op}}$ sub-Gaussian (see, for instance, Vershynin (2012)), hence $\gamma_X\sqrt{\|\Sigma_Z\|_{op}}$ sub-Gaussian as $\|\Sigma_{XZ}\Sigma_i^{-1}\Sigma_{XZ}\|_{op} \leq \|\Sigma_Z\|_{op}$. By using Theorem 5.39 in Vershynin (2012), one has

$$\mathbb{P}\left\{ \frac{1}{n}\|XA^*\|_{op}^2 \leq \|\Sigma_Z\|_{op} + \sqrt{\frac{K}{n}\sqrt{\frac{K}{n}}} \right\} \geq 1 - 2e^{-cK}.$$

The result then follows by $\|\Sigma_Z\|_{op} = O(1)$, $K = O(n)$ and noting that

$$\|XL^*\|_{op}^2 \leq \|XA^*\|_{op}^2 \|B^*\|_{op}^2 \leq \|XA^*\|_{op}^2 \frac{\lambda_1(B^T\Sigma_WB^*)}{\lambda_K(\Sigma_W)}$$

together with part (b) of Assumption 3 and $[\lambda_K(\Sigma_W)]^{-1} = O(1)$. \hfill \Box

C.5 Technical lemmas for controlling the stochastic terms

Recall that $\varepsilon = WB^* + E$ and $P_{\lambda_2}$, $Q_{\lambda_2}$ are defined in (2.4). Further recall that

$$V_\varepsilon = \text{tr}(\Gamma_\varepsilon) = \gamma_w^2\|\Sigma_{W}B^*\|_F^2 + m\gamma_e^2, \quad \Gamma_\varepsilon = \gamma_w^2B^T\Sigma_WB^* + \gamma_e^2I_m.$$  

We first state a lemma which studies the tail behaviour of $\varepsilon$.

Lemma 11. Under Assumption 2, $\varepsilon_{ij}$ is $\gamma_{\varepsilon_{ij}}$ sub-Gaussian for any $i \in [n]$ and $j \in [m]$ with

$$\gamma_{\varepsilon_{ij}}^2 = \gamma_w^2[(B_j^*)^T\Sigma_WB_j^*] + \gamma_e^2.$$  

Furthermore, the random vector $\varepsilon_j$ is $\gamma_{\varepsilon_j}$ sub-Gaussian for any $j \in [m]$ and the random vector $\Gamma_\varepsilon^{-1/2}\varepsilon_i$ is sub-Gaussian with sub-Gaussian constant equal to 1 for any $i \in [n]$.

Proof. Fix any $i \in [n]$ and $j \in [m]$. For any $t \geq 0$, by the independence of $E$ and $W$, we have

$$\mathbb{E}[\exp(t\varepsilon_{ij})] = \mathbb{E}[\exp(tW^T_jB_j^*)] \cdot \mathbb{E}[\exp(tE_{ij})] \leq \exp\left(t^2\gamma_w^2\|\Sigma_{W}B_j^*\|_2^2/2\right) \exp(t^2\gamma_e^2/2)$$

60
Lemma 12. Under Assumption 2, one has

\[
\mathbb{E} \left[ \exp(u^T \Gamma^{-1/2}_\epsilon \varepsilon_i) \right] = \mathbb{E}[\exp(u^T \Gamma^{-1/2}_\epsilon B^* T W_i)] \cdot \mathbb{E}[\exp(u^T \Gamma^{-1/2}_\epsilon E_i)] \\
\leq \exp \left\{ \frac{1}{2} \left[ \gamma_{\epsilon, w}^2 u^T \Gamma^{-1/2}_\epsilon B^* T \Sigma W B^* \Gamma^{-1/2}_\epsilon u + \gamma_{\epsilon, m}^2 u^T \Gamma^{-1/2}_\epsilon I_m \Gamma^{-1/2}_\epsilon u \right] \right\} \\
= \exp(\gamma_{\epsilon, w}^2 \|u\|^2/2).
\]

We used the independence between \( W \) and \( E \) in the first equality and used Assumption 2 to derive the second line. This completes the proof. \( \square \)

We present several lemmas which control different terms related with \( \varepsilon, W \) and \( E \). The following lemma states the deviation inequality of \( \| P_{\lambda_2} \varepsilon \|_F^2, \| P_{\lambda_2}^{1/2} \varepsilon \|_2^2 \) and \( \| \varepsilon^T Q_{\lambda_2} X L^* \|_F \).

**Lemma 12.** Under Assumption 2, one has

\[
\mathbb{P} \left\{ \| P_{\lambda_2} \varepsilon \|_F^2 \leq V_{\epsilon} \left( \sqrt{\text{tr}(P_{\lambda_2}^2)} + \sqrt{2\| P_{\lambda_2} \|_{\text{op}} \log(m/\epsilon)} \right)^2 \right\} \geq 1 - \epsilon \\
\mathbb{P} \left\{ \| P_{\lambda_2}^{1/2} \varepsilon \|_2^2 \leq V_{\epsilon} \left( \sqrt{\text{tr}(P_{\lambda_2})} + \sqrt{2\| P_{\lambda_2} \|_{\text{op}} \log(m/\epsilon)} \right)^2 \right\} \geq 1 - \epsilon \\
\mathbb{P} \left\{ \| \varepsilon^T Q_{\lambda_2} X L^* \|_F \leq \sqrt{m} V_{\epsilon} \sqrt{\| Q_{\lambda_2} \|_{\text{op}} \text{Rem}_2(L^*) \log(m/\epsilon)} \right\} \geq 1 - \epsilon.
\]

**Proof.** We now prove the first result. Note that \( \| P_{\lambda_2} \varepsilon \|_F^2 = \sum_{j=1}^m \| P_{\lambda_2} \varepsilon_j \|_2^2 \). Pick any \( j \in [m] \).

Since \( \varepsilon_j \) is \( \gamma_{\epsilon, j} \) sub-Gaussian from Lemma 11, applying Lemma 17 with \( \gamma = \gamma_{\epsilon, j} \) and \( K = P_{\lambda_2}^2 \) yields

\[
\mathbb{P} \left\{ \varepsilon_j^T P_{\lambda_2} \varepsilon_j > \gamma_{\epsilon, j}^2 \left( \sqrt{\text{tr}(P_{\lambda_2}^2)} + \sqrt{2\| P_{\lambda_2} \|_{\text{op}} t} \right)^2 \right\} \leq e^{-t},
\]

for all \( t \geq 0 \). Since \( \sum_{j=1}^m \gamma_{\epsilon, j}^2 = V_{\epsilon} \), choosing \( t = \log(m/\epsilon) \) and taking the union bounds over \( 1 \leq j \leq m \) complete the proof of the first result. By the same arguments, the second result follows immediately, and we also have

\[
\mathbb{P} \left\{ \| \varepsilon^T Q_{\lambda_2} X L^* \|_F \leq V_{\epsilon} \left( \sqrt{\text{tr}(D)} + \sqrt{2\| D \|_{\text{op}} \log(m/\epsilon)} \right)^2 \right\} \geq 1 - \epsilon
\]

where \( D = Q_{\lambda_2} X L^*(L^*)^T X^T Q_{\lambda_2} \). The third result then follows by observing that

\[
\text{tr}(D) \leq \| Q_{\lambda_2} X L^* \|_F^2 \leq n \| Q_{\lambda_2} \|_{\text{op}} \cdot \text{Rem}_2(L^*)
\]

from (B.5) and \( \| D \|_{\text{op}} \leq \text{tr}(D) \). \( \square \)

The following lemma provides the deviation inequality of \( \max_{1 \leq j \leq p} \| X_j^T Q_{\lambda_2} \varepsilon \|_2 \). Recall that \( M = n^{-1} X^T Q_{\lambda_2}^2 X \) and \( \Gamma_{\epsilon} = \gamma_{\epsilon, w}^2 B^* \Sigma W B^* + \gamma_{\epsilon, m}^2 I_m \).

**Lemma 13.** Under Assumption 2, with probability \( 1 - \epsilon \), one has

\[
\max_{1 \leq j \leq p} \| X_j^T Q_{\lambda_2} \varepsilon \|_2^2 \leq \left( \sqrt{\text{tr}(\Gamma_{\epsilon})} + \sqrt{2\| \Gamma_{\epsilon} \|_{\text{op}} \log(pm/\epsilon)} \right)^2 n \max_{1 \leq j \leq p} M_{jj}.
\]

61
Proof. Fix any $j \in [p]$. Notice that
\[
\|X_j^TQ\lambda_2\varepsilon\|_2^2 = X_j^TQ\lambda_2\varepsilon\Gamma_\varepsilon^{-1/2}\Gamma_\varepsilon^{-1/2}\varepsilon^TQ\lambda_2X_j.
\]
We first show that $\Gamma_\varepsilon^{-1/2}\varepsilon^TQ\lambda_2X_j$ is $\sqrt{nm_{jj}}$ sub-Gaussian. By independence across $i \in [n]$, we have, for any $v \in \mathbb{R}^m$,
\[
\mathbb{E}\left[\exp\left(v^T\Gamma_\varepsilon^{-1/2}\varepsilon^TQ\lambda_2X_j\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(v_i^2TQ\lambda_2X_j\right)\right] \\
\leq \prod_{i=1}^n \exp\left(\|v\|_2^2TQ\lambda_2X_jX_j^TQ\lambda_2\varepsilon_i/2\right) \\
= \exp(\|v\|_2^2nM_{jj}/2).
\]
We used Lemma 11 to derive the second line. Then invoke Lemma 17 with $\gamma_\xi = \sqrt{nm_{jj}}$ and $K = \Gamma_\varepsilon$ gives
\[
P\left\{\|X_j^TQ\lambda_2\varepsilon\|_2^2 > nM_{jj}\left(\sqrt{\text{tr}(\Gamma_\varepsilon)} + \sqrt{2\|\Gamma_\varepsilon\|_{op}}\right)^2\right\} \leq e^{-t}, \quad \text{for all } t \geq 0.
\]
Choose $t = \log(p/\epsilon)$ and take the union bounds over $j \in [p]$ to complete the proof.

The next tail inequality is for $\max_{1 \leq j \leq p} \|X_j^TE\|_2$, derived based on the quadratic form of a sub-Gaussian random vector.

**Lemma 14.** Under Assumption 2, with probability $1 - \epsilon$, one has
\[
\max_{1 \leq j \leq p} \|X_j^TE\|_2^2 \leq \gamma_\varepsilon^2 \left(\sqrt{m} + \sqrt{2\log(p/\epsilon)}\right)^2 n \max_{1 \leq j \leq p} \hat{\Sigma}_{jj}.
\]

**Proof.** Pick any $j \in [p]$. Note that $E^TX_j$ is $\gamma_\varepsilon\sqrt{n\hat{\Sigma}_{jj}}$ sub-Gaussian. Indeed, for any $u \in \mathbb{R}^m$, we have
\[
\mathbb{E}[\exp(\langle u, E^TX_j \rangle)] = \prod_{t=1}^n \prod_{i=1}^m \mathbb{E}\left[\exp(u_iX_{ij}E_{ti})\right] \\
\leq \prod_{t=1}^n \prod_{i=1}^m \exp(u_i^2X_{ij}^2\gamma_\varepsilon^2/2) \\
= \exp(\|u\|_2^2X_j^TE_j\gamma_\varepsilon^2/2)
\]
by using the independence of entries of $E$ and $E_{ti}$ is $\gamma_\varepsilon$ sub-Gaussian. Applying Lemma 17 with $K = I_m$ and $\gamma_\xi = \gamma_\varepsilon\sqrt{n\hat{\Sigma}_{jj}}$ gives
\[
P\left\{X_j^TE^TX_j > \gamma_\varepsilon^2n\hat{\Sigma}_{jj}\left(\sqrt{m} + \sqrt{2t}\right)^2\right\} \leq e^{-t}.
\]
Choosing $t = \log(p/\epsilon)$ and taking the union bounds over $1 \leq j \leq p$ conclude the proof.

The following lemma states the tail behaviour of the operator norm of $\Sigma_W^{-1/2}W^TW\Sigma_W^{-1/2}$. 62
Lemma 15. Under Assumption 2, with probability \(1 - 2e^{-cK}\), one has
\[
\frac{1}{n} \left\| \Sigma W^{-1/2} W^T W \Sigma W^{-1/2} \right\|_{op} \leq 1 + C \left( \sqrt{\frac{K}{n}} \vee \frac{K}{n} \right)
\]
where \(c = c(\gamma_w)\) and \(C = C(\gamma_w)\) are positive constants.

Proof. The result follows directly from the proof of Theorem 5.39 in Vershynin (2012).

The following lemma states the deviation inequality of \(\|n^{-1} \xi - \Sigma \xi\|_F\).

Lemma 16. Under Assumption 2, one has
\[
P \left\{ \left\| \frac{1}{n} \xi^T \xi - \Sigma \xi \right\|_F \leq cV \xi \left( \sqrt{\log \left( \frac{m}{n} \vee \frac{m}{n} \right)} + \sqrt{2 \|K\|_{op} t} \right) \right\} \geq 1 - 2m^{-c'}
\]
for some absolute constants \(c, c' > 0\).

Proof. Fix any \(j, \ell \in [m]\). We first upper bound \(|n^{-1} \xi_j^T \xi_\ell - (\Sigma \xi)_j \ell|\). Since entries of \(\xi_j\) and \(\xi_\ell\) are \(\gamma_{\xi_j}\) and \(\gamma_{\xi_\ell}\) sub-Gaussian, respectively, from Lemma 11, invoking Lemma 15 in Bing et al. (2019) with \(t = \min \{ \sqrt{\log(m)/n} \vee \log(m)/n \} \) gives
\[
P \left\{ \left| \frac{1}{n} \xi_j^T \xi_\ell - (\Sigma \xi)_j \ell \right| \leq c' \gamma_{\xi_j} \gamma_{\xi_\ell} \left( \sqrt{\log(m)/n} \vee \log(m)/n \right) \right\} \geq 1 - 2m^{-c'}
\]
for some absolute constants \(c, c' > 0\). The result then follows by taking the union bounds over \(1 \leq j, \ell \leq m\) and noting that \(\sum_j \sum_\ell \gamma_{\xi_j}^2 \gamma_{\xi_\ell}^2 = V_\xi^2\).

C.6 An algebraic fact and one auxiliary lemma

We first state an algebraic fact that is used in our analysis.

Fact 1. Let \(Q_{\lambda_2}\) be defined in (2.4). Then
\[
Q_{\lambda_2} X = \lambda_2 X (\hat{\Sigma} + \lambda_2 I_p)^{-1}.
\]

Proof. The proof follows by noting that
\[
Q_{\lambda_2} X = X - X (X^T X + n \lambda_2 I_p)^{-1} X^T X = n \lambda_2 X (X^T X + n \lambda_2 I_p)^{-1}
\]
and the definition \(\hat{\Sigma} = X^T X / n\).

The following lemma is used in our analysis. The tail inequality is for a quadratic form of sub-Gaussian random vector. It is a slightly simplified version of Lemma 8 in Hsu et al. (2014).

Lemma 17. Let \(\xi \in \mathbb{R}^d\) be a \(\gamma_{\xi}\) sub-Gaussian random vector. For all symmetric positive semidefinite matrices \(K\), and all \(t \geq 0\),
\[
P \left\{ \xi^T K \xi > \gamma_{\xi}^2 \left( \sqrt{\text{tr}(K)} + \sqrt{2\|K\|_{op} t} \right)^2 \right\} \leq e^{-t}.
\]

Proof. From Lemma 8 in Hsu et al. (2014), one has
\[
P \left\{ \xi^T K \xi > \gamma_{\xi}^2 \left( \text{tr}(K) + 2\sqrt{\text{tr}(K^2)} t + 2\|K\|_{op} t \right) \right\} \leq e^{-t},
\]
for all \(t \geq 0\). The result then follows from \(\text{tr}(K^2) \leq \|K\|_{op} \text{tr}(K)\).
D Comparison of the multivariate ridge estimation and the reduced-rank estimation

In this section, we compare the multivariate ridge regression with the reduced-rank estimator under model $Y = X_L + W B + E$ when $\Theta = 0$. Comparing to the commonly studied low-rank regression in the literature, the noise level here is much larger since it follows a factor structure with diverging eigenvalues.

In the following we will demonstrate the advantage of using the multivariate ridge regression in (2.3) with $\lambda_1 = 0$ over the reduced-rank estimator. We first show that the ridge-type estimator has smaller PMSE than the reduced-rank estimator when $\|L\|_F$ is small or moderate. Second, we further show that the ridge-type estimator is more robust to the noise level than the reduced-rank estimator.

We follow the data generating process in simulation studies and choose $n = 80$, $p = 120$, $m = 30$ and $\rho = 0.3$. To change the strength of $\|L\|_F$, we vary $\eta \in \{0.05, 0.1, 0.15, \ldots, 0.5, 0.55\}$. For each $\eta$, we randomly generate 100 datasets and the averaged PMSEs of $\hat{L}^{(RR)}$ (RR) and $\hat{L}^{(Ridge)}$ (Ridge) are shown in Figure 7. Note that we provide the true $K$ for $\hat{L}^{(RR)}$. As seen in the first panel, Ridge has much smaller PMSE than RR when $\eta$ is small. Moreover, it seems that RR does not provide consistent prediction when the noise has diverging eigenvalues.

To show the robustness to the noise level, we use the same setting and fix $\eta = 0.2$. Recall that $W_{ik} \sim N(0, \sigma_W^2 = 1)$. We then change the noise level by varying $\sigma_W^2 \in \{0.6, 0.8, 1, \ldots, 2.8, 3\}$. The averaged PMSEs of RR and Ridge are shown in the second panel of Figure 7. It is easy to see that Ridge is much more robust to the magnitude of the noise level and outperforms RR by a large margin.

![Figure 7: The averaged PMSEs of RR and Ridge when we vary $\eta$ and $\sigma_W$ separately.](image-url)