Submitted to the Annals of Statistics

ADAPTIVE ESTIMATION OF MULTIVARIATE REGRESSION WITH HIDDEN VARIABLES

By Xin Bing and Yang Ning and Yaosheng Xu

Cornell University

A prominent concern of scientific investigators is the presence of unobserved hidden variables in association analysis. Ignoring hidden variables often yields biased statistical results and misleading scientific conclusions. Motivated by this practical issue, this paper studies the multivariate regression with hidden variables, $Y = (\Psi^*)^T X + (B^*)^T Z + E$, where $Y \in \mathbb{R}^m$ is the response vector, $X \in \mathbb{R}^p$ is the observable feature, $Z \in \mathbb{R}^K$ represents the 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.

Though Ψ^* is shown to be non-identifiable due to the presence of hidden variables, we propose to identify the projection of Ψ^* onto the orthogonal complement of the row space of B^*, denoted by Θ^*. The quantity $(\Theta^*)^T X$ measures the effect of X on Y that cannot be explained through the hidden variables, and thus Θ^* is treated as the parameter of interest. Motivated by the identifiability proof, we propose a novel estimation algorithm for Θ^*, called HIVE, under homoscedastic errors. The first step of the algorithm estimates the best linear prediction of Y given X, in which the unknown coefficient matrix exhibits an additive decomposition of Ψ^* and a dense matrix due to the correlation between X and Z. Under the sparsity assumption on Ψ^*, we propose to minimize a penalized least squares loss by regularizing Ψ^* and the dense matrix via group-lasso and multivariate ridge, respectively. Non-asymptotic deviation bounds of the in-sample prediction error are established. Our second step estimates the row space of B^* by leveraging the covariance structure of the residual vector from the first step. In the last step, we estimate Θ^* via projecting Y onto the orthogonal complement of the estimated row space of B^* to remove the effect of hidden variables. Non-asymptotic error bounds of our final estimator of Θ^*, 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 is adaptive to the unknown sparsity of Θ^* induced by the sparsity of Ψ^*. The model identifiability, estimation algorithm 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

Primary 62H12, 62J07

Keywords and phrases: High-dimensional models, multivariate regression, shrinkage, non-sparse estimation, hidden variables, confounding, surrogate variable analysis

results.

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 [3]. Most of the existing research on multivariate regression assumes that the collected predictors are sufficient to explain the responses. However, due to cost constraints or ethical issues, oftentimes there exist unmeasured hidden variables that are associated with the responses as well. 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

$$Y = (\Psi^*)^T X + (B^*)^T Z + E, \tag{1.1}$$

where $\Psi^* \in \mathbb{R}^{p\times m}$ 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 Σ_E. The random error E is independent of (X, Z) and the hidden variable Z is possibly correlated with X. The number of hidden variables, K, is unknown and assumed to be less than m. As we can subtract means from both sides of (1.1), we consider Y, X and Z have mean zero. Without loss of generality, we assume $\Sigma = \mathrm{Cov}(X)$ and $\Sigma_Z = \mathrm{Cov}(Z)$ are strictly positive definite and $\mathrm{rank}(B^*) = K$. 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 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and a response matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$. 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 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 unobserved confounders, surrogate variable analysis (SVA) based on model (1.1) has been proposed for the analysis of biological data [38, 39, 45, 19, 28, 31, 44]. 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 [37, 47, 41]. Since p is small, the existing

SVA methods apply the ordinary least squares (OLS) $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ to estimate the main regression effect and then remove the bias of the OLS estimator, originated from the correlation between Z and X. However, to avoid confounding issues, researchers tend to collect as many features as possible and then adjust them in the regression model. In this case, p can be large and even much larger than n, whence the existing SVA methods are not applicable as the OLS estimator 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 be also framed as linear structural equation models [32]. 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 $(X, Z) \to Y$ 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 observable, model (1.1) and (1.2) can be viewed as linear structural equation models with hidden variables [22]. Using the terminology in causal mediation analysis, the parameter Ψ^* in (1.1) represents the direct causal effect of X on Y. 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 with the independence between X and W.

1.1. *Our contributions.*

Identifiability. Our first contribution is to investigate the identifiability of model (1.1). We show that Ψ^* in model (1.1) is not identifiable in Proposition 1 of Section 2.1. This motivates us to focus on an alternative estimand, the projection of Ψ^* onto the orthogonal complement of row space of B^*, which is identifiable and has desirable interpretations.

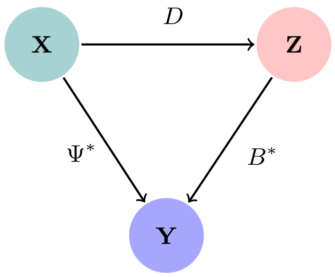

Fig 1: Illustration of the DAG under model (1.1) and (1.2)

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 $W = Z - (A^*)^T X$ its residual, where

$$A^* = \left\{\mathbb{E}\left[XX^T\right]\right\}^{-1} \mathbb{E}\left[XZ^T\right] \in \mathbb{R}^{p\times K}. \tag{1.3}$$

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

$$\begin{aligned} Y &= (\Psi^* + A^* B^*)^T X + (B^*)^T W + E \\ &= (\underbrace{\Psi^* P_{B^*}^{\perp}}_{\Theta^*} + \Psi^* P_{B^*} + \underbrace{A^* B^*}_{L^*})^T X + \underbrace{(B^*)^T W + E}_{\varepsilon}, \end{aligned} \tag{1.4}$$

where $P_{B^*} = B^{*T}(B^* B^{*T})^{-1} B^* \in \mathbb{R}^{m\times m}$ is the projection matrix onto the row space of B^*, $P_{B^*}^{\perp} = \boldsymbol{I}_m - P_{B^*}$, $L^* = A^* B^*$ and the residual vector $\varepsilon = (B^*)^T W + E$ satisfies $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Cov}(X, \varepsilon) = 0$.

In (1.4), we decompose Ψ^* into two components $\Psi^* P_{B^*}^{\perp}$ and $\Psi^* P_{B^*}$. The former denoted by Θ^* is the projection of Ψ^* onto the orthogonal complement of row space of B^*. Since ${\Theta^*}^T X$ is orthogonal to ${B^*}^T Z$, ${\Theta^*}^T X$ measures the effect of X on Y in the multivariate regression (1.1) that cannot be explained through the hidden variable effect. In the mediation analysis, we can refer to Θ^* as "partial" direct effect of X on Y.

In Propositions 2 and 3 of Section 2.1 we establish the sufficient and necessary conditions for the identifiability of Θ^* when the error E is homoscedastic, that is $\Sigma_E = \tau^2 \boldsymbol{I}_m$. This covariance structure of the residual vector ε is crucial to show the results, as detailed in Section 2.1. However, the sufficiency no longer holds in the presence of heteroscedastic error, when Σ_E is a diagonal matrix with unequal entries. Inspired by [50], 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 Θ^*. We show in Proposition 9 of Section 4 that this incoherence condition guarantees the identifiability of Θ^* in the heteroscedastic case.

Summarizing, the parameter Θ^* is identifiable under both homoscedastic and heteroscedastic errors. We focus on the estimation of Θ^* throughout the paper. It is worth mentioning that our analysis of Θ^* carries over to Ψ^*, under $\Psi^* P_{B^*} \to 0$ as $m = m(n) \to \infty$, a common condition of Ψ^* in the SVA literature [37]. Indeed, even if Ψ^* is of primary interest, the information in Θ^* may still be helpful to infer Ψ^*; see Section 5.5 for details.

Estimation of Θ^*. Our second contribution is to propose a new method for estimating Θ^*. 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 Θ^*, 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, [11, 14, 40, 42, 49], just to name a few. While Ψ^* is not identifiable, we assume Ψ^* lies in the space Ω

$$\Psi^* \in \Omega := \left\{ M \in \mathbb{R}^{p\times m} : \|M\|_{\ell_0/\ell_2} \le s_* \right\}, \tag{1.5}$$

where $s_* \le p$ and $\|M\|_{\ell_0/\ell_2} = \sum_{j=1}^p 1_{\{\|M_{j\cdot}\|_2 \neq 0\}}$ is the number of nonzero rows. As a result, (1.5) implies that $\Theta^* \in \Omega$ as

$$\|\Theta^*\|_{\ell_0/\ell_2} = \|\Psi^* P_{B^*}^{\perp}\|_{\ell_0/\ell_2} \le s_*. \tag{1.6}$$

Our estimation procedure consists of three steps: first estimate the best linear prediction of Y given X; then estimate the row space of B^* and finally estimate Θ^*.

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 [49] and the multivariate ridge penalty. The group-lasso penalty aims to exploit the row-wise sparsity of Ψ^* 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 Section 2.2.1 for detailed discussions of computational and theoretical advantages of our estimator over other competing methods.

Our second step is to estimate the row space of B^* or equivalently P_{B^*}. When the noise is homoscedastic, we directly apply the principle component analysis (PCA) to the sample covariance matrix of the estimated residuals (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 [50] to our setting.

In the last step, we project Y onto the orthogonal complement of the estimated row space of B^* to remove the effect of hidden variables, and then recover Θ^* by applying group-lasso to the projected Y.

Our entire procedure is summarized in Algorithm 1, called HIVE, representing $\underline{\text{HI}}$dden $\underline{\text{V}}$ariable adjustment $\underline{\text{E}}$stimation. 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 on practical implementations in Section 5, including estimation of the number of hidden variables (K), the consequence of overestimating/underestimating K, the choice of tuning parameters, data standardization and practical usage of Θ^* for inferring Ψ^*.

Statistical guarantees. Our third contribution is to establish theoretical properties of our procedure. In Theorem 4 of Section 4, 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 components: bias and variance terms 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 $\Psi^* = 0$. Thus, the rate of our estimator matches the best possible rate even if $L^* = 0$ or $\Psi^* = 0$ were known a priori.

We further provide theoretical guarantees for the estimation of Θ^*. In particular, we establish in Theorem 6 a general non-asymptotic upper bound of the estimation error of our estimator $\widetilde{\Theta}$ based on any estimator $\widehat{P}$ of P_{B^*}. As expected, the estimation error of $\widetilde{\Theta}$ depends on how accurately $\widehat{P}$ estimates P_{B^*}. When P_{B^*} can be estimated accurately enough, our estimator $\widetilde{\Theta}$ achieves the optimal rate in the oracle case with known B^* (see the subsequent paragraph of Theorem 6). However, if the estimation error of P_{B^*} is relatively large, we can balance this term with the error of the group-lasso to attain a more refined rate via a suitable choice of the regularization parameter. In Theorem 7 of Section 3.2 and Theorem 10 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 6 provide the final upper bounds of the estimation error of $\widetilde{\Theta}$. To deal with 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 essential to the proof of Theorem 10 and can be of its own interest.

1.2. *Related literature.* This work is most related to the literature on surrogate variable analysis (SVA) in which the parameter of interest is Ψ^*. Though our target $\Theta^* = \Psi^* P_{B^*}^{\perp}$ is different from Ψ^*, our analysis of Θ^* is applicable to Ψ^* under the condition $\Psi^* P_{B^*} = 0$ (or $\Psi^* P_{B^*} \to 0$). We thus compare our results under $\Psi^* P_{B^*} = 0$ with the existing SVA literature. For the identifiability of Ψ^*, [28, 47] assumed that there exists a *known* subset $J \in \{1, \ldots, m\}$ such that the $p \times |J|$ submatrix $\Psi^*_J = 0$. 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 [47, 41] assumes that each row $\Psi^*_{j\cdot} \in \mathbb{R}^m$ is sparse with $\|\Psi^*_{j\cdot}\|_0 \leq (m-a)/2$ for some $a > K$ and any $K \times a$ submatrix of B^* is of rank K. This assumption rules out the possibility that B^* could be sparse and the resulting sparsity pattern of Ψ^* differs from (1.5), considered in this work. [37] assumed the condition $\Psi^* P_{B^*} \to 0$ as $m \to \infty$. In contrast, our identifiability result of Θ^* (which is Ψ^* under $\Psi^* P_{B^*} = 0$) 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.

[20] 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 [33, 17] 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, the parameter of interest is Θ^*, and thus we do not need the identifiability assumptions in [20]. 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 ith canonical basis vector of $\mathbb{R}^p$. The identifiability assumption in [20] does not hold, because the low rank matrix $L^* = A^* B^* = e_i e_i^T$ is too sparse and cannot be distinguished from the sparse matrix Ψ^*. How-

ever, our target $\Theta^* \in \{\Psi^*(\boldsymbol{I}_p - e_i e_i^T) : \Psi^* \in \mathbb{R}^{p\times p}\}$ is still identifiable 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 [20]. Second, due to the differences of identifiability, our estimation algorithm (HIVE or H-HIVE) 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 Appendix D.

Recently, [22] 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 [22] is adapted from the SVA literature, and therefore has the same drawback as SVA. In another recent paper, [18] 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 β, they assumed that $\|b\|_2$ is sufficiently small. Their estimation method generalizes the lava estimator [21]. Unlike this work, we consider a different setting where the response Y is multivariate and, consequently, both our identifiability and estimation procedures (HIVE and H-HIVE) are completely different from theirs. Our theoretical results in Corollary 8 and its subsequent Remark 6 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.

1.3. *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 Ψ^*+L^* 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 Θ^* 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 Θ^* 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, data standardization and practical usage of Θ^* for inferring Ψ^*. Simulation results and real data applications are presented in Sections 6 and 7. All proofs and supplementary simulation results are deferred to Supplement to "Adaptive Estimation of Multivariate Regression with Hidden Variables".

1.4. *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 ℓ_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_{IJ} as the $|I| \times |J|$ submatrix of M with row and column indices corresponding to I and J, respectively. In particular, $M_{I\cdot}$ denotes the $|I| \times d_2$ submatrix and M_J denotes the $d_1 \times |J|$ submatrix. Further write $\|M\|_{\ell_p/\ell_q} = (\sum_{j=1}^{d_1} \|M_{j\cdot}\|_{\ell_q}^p)^{1/p}$ and denote by $\|M\|_{\ell_0}$, $\|M\|_{op}$, $\|M\|_F$ and $\|M\|_\infty$, respectively, the element-wise ℓ_0 norm, the operator norm, the Frobenius norm and the element-wise sup-norm of M. For any symmetric matrix M, we write $\lambda_k(M)$ for its kth 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 $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Throughout the paper, we will write $\widehat{\Sigma} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$ with non-zero eigenvalues $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$ and $q := \operatorname{rank}(\boldsymbol{X})$.

2. Identifiability and estimation under homoscedastic noise.

As seen in the Introduction, Ψ^* is not identifiable under model (1.1) due to the presence of hidden variables. The following proposition formally shows this.

PROPOSITION 1. *Under model (1.1), or equivalently (1.4), suppose $Z \in \mathbb{R}^K$ has continuous support and $A^* \neq 0$. Then Ψ^* is not identifiable.*

Despite of the non-identifiability of Ψ^*, we proceed to show that $\Theta^* = \Psi^* P_{B^*}^{\perp}$ is identifiable when the error E is homoscedastic. Our identifiability procedure is constructive and leads to a computationally efficient estimation algorithm for Θ^*.

2.1. *Identifiability of* Θ^*.

We start by describing our procedure of identifying Θ^* from model (1.4) in three steps:

(1) identify the coefficient matrix

$$F^* := \Psi^* + A^* B^* = \Psi^* + L^*;$$

(2) identify $\Sigma_\varepsilon := \mathrm{Cov}(\varepsilon)$ with $\varepsilon = (B^*)^T W + E$ and use it to construct P_{B^*}, the projection matrix onto the row space of B^*;
(3) identify Θ^* from $(\boldsymbol{I}_m - P_{B^*})Y$.

Recall that $W = Z-(A^*)^T X$ is independent of E with A^* defined in (1.3). In step (2), a key observation from model (1.4) is that, under homoscedastic error, the covariance matrix of $\varepsilon = (B^*)^T W + E$ satisfies

$$\Sigma_\varepsilon = (B^*)^T \Sigma_W B^* + \Sigma_E = (B^*)^T \Sigma_W B^* + \tau^2 \boldsymbol{I}_m, \tag{2.1}$$

where $\Sigma_W = \mathrm{Cov}(W)$. Recall that $\mathrm{rank}(B^*) = K < m$. Provided that Σ_W has full rank, (2.1) implies that the row space of B^* coincides with the space spanned by the eigenvectors of Σ_ε corresponding to its largest K eigenvalues. We thus propose to identify P_{B^*} via the eigenspace of Σ_ε. Step (3) uses

$$\begin{aligned} P_{B^*}^{\perp} Y &= (\Psi^* P_{B^*}^{\perp})^T X + (B^* P_{B^*}^{\perp})^T Z + P_{B^*}^{\perp} E \\ &= (\Theta^*)^T X + P_{B^*}^{\perp} E, \end{aligned} \tag{2.2}$$

where $P_{B^*}^{\perp} = \boldsymbol{I}_m - P_{B^*}$. This further implies

$$\Theta^* = \left[\mathrm{Cov}(X)\right]^{-1} \mathrm{Cov}\left(X, P_{B^*}^{\perp} Y\right).$$

The following proposition summarizes the identifiability of Θ^* under the homoscedastic error.

Proposition 2. *Under model (1.4), Θ^* is identifiable if either of the following holds:*

(1) $\Psi^* P_{B^*} + A^* B^* = 0$;
(2) $\mathrm{rank}(\Sigma_W) = K$ and $\Sigma_E = \tau^2 \boldsymbol{I}_m$.

Case (1) implies that $\Psi^* P_{B^*}$, the direct effect of X on Y explained by Z, can be exactly offset by the indirect effect $A^* B^*$. In this case, Θ^* can be recovered by regressing Y on X directly. Since this is rarely the case in practice, we will focus on $\Psi^* P_{B^*} + A^* B^* \neq 0$. Case (2) requires $\mathrm{rank}(\Sigma_W) = K$ in addition to $\Sigma_E = \tau^2 \boldsymbol{I}_m$. We show, in Proposition 3 below, that $\mathrm{rank}(\Sigma_W) = K$ is also necessary for identifying Θ^* if $\mathbb{E}[W|X] = 0$ and $\Psi^* P_{B^*} + A^* B^* \neq 0$.

Proposition 3. *Under model (1.4) with $\mathbb{E}[W|X] = 0$, assume $\Psi^* P_{B^*} + A^* B^* \neq 0$. If $\mathrm{rank}(\Sigma_W) < K$, then Θ^* is not identifiable.*

Combining Propositions 2 and 3 concludes that, under homoscedastic error, $\Psi^* P_{B^*} + A^* B^* \neq 0$ and $\mathbb{E}[W|X] = 0$, Θ^* is identifiable if and only if

$\operatorname{rank}(\Sigma_W) = K$. The condition $\mathbb{E}[W|X] = 0$ is satisfied in many interesting scenarios, such as the structured equation model (1.2) and the multivariate Gaussian model for (Z, X). In practice, recalling that $W = Z - A^{*T}X$, $\operatorname{rank}(\Sigma_W) = K$ is 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. Therefore, throughout the paper, we assume $\operatorname{rank}(\Sigma_W) = K$.

Remark 1. In the SVA literature, [37] assumed $\Psi^* P_{B^*} \to 0$ as $m = m(n) \to \infty$. Under this condition, we obtain $\Theta^* \approx \Psi^*$ for sufficiently large m. In this case, Propositions 2 and 3 provide sufficient and necessary conditions for the identifiability of Ψ^* for m large enough. Therefore, our analysis provides complete identifiability results for SVA and it further generalizes to the setting that $\Psi^* P_{B^*} \not\to 0$.

Remark 2. Our identifiability results are established when the rows of $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ are viewed as i.i.d. realizations of a random vector $X \in \mathbb{R}^p$. When $\boldsymbol{X}$ is treated as a fixed design matrix, Proposition 1 in [41] provides sufficient conditions on Ψ^* and other quantities (such as the sparsity of rows of Ψ^* and the magnitude of B^*) under which Ψ^* becomes identifiable for n sufficiently large.

2.2. *Estimation of* Θ^*.

Given the data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, our estimation procedure follows the same steps in the analysis of the model identifiability: (1) first estimate $\boldsymbol{X}F^*$; (2) then estimate Σ_ε and P_{B^*}; (3) finally estimate Θ^*.

2.2.1. *Estimation of* XF^*.

Recall that $F^* = \Psi^* + L^*$ is identifiable, where $L^* = A^*B^*$ is a dense matrix and Ψ^* is a row-wise sparse matrix satisfying (1.5). We propose to estimate F^* by $\widehat{F} = \widehat{\Psi} + \widehat{L}$ where $\widehat{\Psi}$ and $\widehat{L}$ are obtained by solving the following optimization problem

$$(\widehat{\Psi},\ \widehat{L}) = \arg\min_{\Psi, L} \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}(\Psi + L)\|_F^2 + \lambda_1 \|\Psi\|_{\ell_1/\ell_2} + \lambda_2 \|L\|_F^2 \tag{2.3}$$

with some tuning parameters $\lambda_1, \lambda_2 \geq 0$. Our estimator is designed to recover both the sparse matrix Ψ^* via the group-lasso regularization [49] and the dense matrix L^* via the multivariate ridge regularization. Since our goal in this step is to estimate the best linear predictor $\boldsymbol{X}F^*$, there is no need to separate Ψ^* from L^*. 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 $(\widehat{\Psi}, \widehat{L})$ be any solution of (2.3), and denote*

$$P_{\lambda_2} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + n\lambda_2 \boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T, \qquad Q_{\lambda_2} = \boldsymbol{I}_n - P_{\lambda_2} \tag{2.4}$$

for any $\lambda_2 \geq 0$ such that P_{λ_2} exists. Then $\widehat{\Psi}$ is the solution of the following problem

$$\widehat{\Psi} = \arg\min_{\Psi} \frac{1}{n}\left\| Q_{\lambda_2}^{1/2}(\boldsymbol{Y} - \boldsymbol{X}\Psi)\right\|_F^2 + \lambda_1 \|\Psi\|_{\ell_1/\ell_2}, \tag{2.5}$$

and $\widehat{L} = (\boldsymbol{X}^T\boldsymbol{X} + n\lambda_2\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\Psi})$, where $Q_{\lambda_2}^{1/2}$ is the principal matrix square root of Q_{λ_2}. Moreover, we have

$$\boldsymbol{X}\widehat{F} = \boldsymbol{X}(\widehat{\Psi} + \widehat{L}) = P_{\lambda_2}\boldsymbol{Y} + Q_{\lambda_2}\boldsymbol{X}\widehat{\Psi}. \tag{2.6}$$

Lemma 1 characterizes the role of the regularization parameters λ_2 and λ_1. When $\lambda_2 \to 0$, we have $P_{\lambda_2} \approx \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T$ and $Q_{\lambda_2}\boldsymbol{X} \approx 0$ with $(\boldsymbol{X}^T\boldsymbol{X})^+$ being the Moore-Penrose inverse of $\boldsymbol{X}^T\boldsymbol{X}$. Thus, $\boldsymbol{X}\widehat{F} \approx \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^+\boldsymbol{X}^T\boldsymbol{Y}$ and $\widehat{F}$ reduces to the minimum norm ordinary least squares estimator (see, for instance, [15]). On the other hand, when $\lambda_2 \to \infty$, we have $Q_{\lambda_2} \approx \boldsymbol{I}_n$ and $\widehat{L} \approx 0$ whence $\widehat{F} \approx \widehat{\Psi}$ essentially becomes the group-lasso estimator. Later in Remark 4 of Section 3.1, we will take a closer look at this phenomenon in terms of the convergence rates of $\|\boldsymbol{X}\widehat{F} - \boldsymbol{X}F^*\|_F$ under the orthogonal design. The tuning parameter λ_1 only appears in (2.5) and its magnitude controls the sparsity level of the group-lasso estimator $\widehat{\Psi}$. Lemma 1 also implies that the estimator $(\widehat{\Psi}, \widehat{L})$ 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 $(\widehat{\Psi}, \widehat{L})$ to be any of the solutions and the resulting best linear predictor $\boldsymbol{X}\widehat{F} = \boldsymbol{X}(\widehat{\Psi} + \widehat{L})$ satisfies the desired deviation bounds stated in Theorem 4 of Section 3.1.

In applications when both m and p are large while K is small, L^* can be also viewed as a low-rank matrix with rank K. One common approach of estimating a low-rank matrix is to either impose a rank constraint on the matrix known as the reduced-rank approach [35] or regularize its nuclear norm. We emphasize that, under model (1.4), our approach with the ridge penalty has both theoretical and computational advantages over these two methods. We defer to Appendix D for both theoretical and numerical comparisons.

Finally, we comment that our method (2.3) can be viewed as the multivariate generalization of the lava approach proposed by [21]; see also [18]. Lava estimates the sum of a sparse vector β and a dense vector b in linear regression problem $\boldsymbol{y} = \boldsymbol{X}(\beta + b) + \boldsymbol{\epsilon}$ by minimizing the least squares loss plus the penalty $\lambda_1\|\beta\|_1 + \lambda_2\|b\|_2^2$. As explained in [21], lava is intrinsically different from the elastic net as lava penalizes both β and b and the estimate of $(\beta+b)$ is non-sparse, whereas elastic net uses the penalty $\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2$ and typically yields a sparse estimate of β. 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 $B^* = VDU^T$, 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 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 Σ_ε up to an orthogonal matrix. We thus propose to first estimate Σ_ε by

$$\widehat{\Sigma}_\varepsilon = \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{F}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{F}\right) \tag{2.7}$$

with $\widehat{F}$ obtained from (2.3) and then estimate P_{B^*} by $\widehat{P}_{B^*} = \widehat{U}\widehat{U}^T$, where $\widehat{U}$ consists of the eigenvectors of $\widehat{\Sigma}_\varepsilon$ 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 Θ^*.* After estimating P_{B^*} by $\widehat{P}_{B^*}$, motivated by (2.2), we propose to estimate Θ^* by

$$\widetilde{\Theta} = \arg\min_{\Theta} \frac{1}{n}\left\|\boldsymbol{Y}\left(\boldsymbol{I}_m - \widehat{P}_{B^*}\right) - \boldsymbol{X}\Theta\right\|_F^2 + \lambda_3\|\Theta\|_{\ell_1/\ell_2} \tag{2.8}$$

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 $\boldsymbol{Y}(\boldsymbol{I}_m - \widehat{P}_{B^*})$.

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

Algorithm 1 The HIVE procedure for estimating Θ^*.

```
Require: Data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, rank $K$, tuning parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$.
Estimate $\boldsymbol{X}\widehat{F}$ with $\widehat{F} = \widehat{\Psi} + \widehat{L}$ by solving (2.3).
Obtain $\widehat{\Sigma}_\varepsilon$ from (2.7).
Compute $\widehat{P}_{B^*} = \widehat{U}\widehat{U}^T$ where $\widehat{U}$ are the first $K$ eigenvectors of $\widehat{\Sigma}_\varepsilon$.
Estimate $\Theta^*$ by $\widetilde{\Theta}$ 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 $\boldsymbol{X}$ is considered to be deterministic and the analysis can be done similarly for random design by first conditioning on $\boldsymbol{X}$. Recall from model (1.4) that W is uncorrelated with X. To simplify the analysis under the fixed design scenario, we assume the independence between X and W in order to derive the deviation bounds of their cross product. We expect that the same theoretical guarantees hold under $\mathrm{Cov}(X, W) = 0$ by using more tedious arguments. We start with the following assumptions on the error matrices $\boldsymbol{W} \in \mathbb{R}^{n\times K}$ and $\boldsymbol{E} \in \mathbb{R}^{n\times m}$.

Assumption 1. *Let γ_w and γ_e denote some positive constants.*

(1) Assume $\{\Sigma_W^{-1/2}\boldsymbol{W}_{i\cdot}\}_{i=1}^n$ are i.i.d. γ_w sub-Gaussian random vectors[1]*, where $\Sigma_W = Cov(\boldsymbol{W}_{i\cdot})$.*

(2) For any fixed $1 \le j \le m$, $\{\boldsymbol{E}_{ij}\}_{i=1}^n$ are i.i.d. γ_e sub-Gaussian[2]*. For any fixed $1 \le i \le n$, $\{\boldsymbol{E}_{ij}\}_{j=1}^m$ are independent.*

Since part (2) of Assumption 1 does not assume $\boldsymbol{E}_{ij}$ are identically distributed across $1 \le j \le m$, this assumption is applicable to both homoscedastic and heteroscedastic errors, provided that $\max_{1\le j\le p} \mathrm{Var}(\boldsymbol{E}_{ij}) \le \gamma_e^2$. We assume $\Sigma_E = \mathrm{Cov}(E) = \tau^2\boldsymbol{I}_m$ throughout this section and defer the dicussion of the heteroscedastic case to Section 4. Finally, recall from (1.5) that $\|\Psi^*\|_{\ell_0/\ell_2} \le s_*$.

3.1. *Statistical guarantees of estimating XF^*.*

To establish theoretical properties for $\boldsymbol{X}\widehat{F}$ obtained from (2.5), we first generalize the design impact factor of $\boldsymbol{X}$ in [21] to multivariate regression settings. Denote $\widetilde{\boldsymbol{X}} = Q_{\lambda_2}^{1/2}\boldsymbol{X}$, where Q_{λ_2} is defined in (2.4). For notational simplicity, we suppress the dependence of $\widetilde{\boldsymbol{X}}$ on λ_2. For any constant $c > 0$ and matrix $\Psi_0 \in \mathbb{R}^{p\times m}$, define the design impact factor as

$$\kappa_1(c, \Psi_0, \lambda_1, \lambda_2) := \inf_{\Delta\in\mathcal{R}(c,\Psi_0,\lambda_1,\lambda_2)} \frac{\|\widetilde{\boldsymbol{X}}\Delta\|_F/\sqrt{n}}{\|\Psi_0\|_{\ell_1/\ell_2} - \|\Psi_0+\Delta\|_{\ell_1/\ell_2} + c\|\Delta\|_{\ell_1/\ell_2}}, \tag{3.1}$$

[1] A random vector X is γ sub-Gaussian if $\langle u, X\rangle$ is γ sub-Gaussian for any $\|u\|_2 = 1$.

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

where

$$\mathcal{R}(c,\Psi_0,\lambda_1,\lambda_2)=\Big\{\Delta\in\mathbb{R}^{p\times m}\setminus\{0\}:\\ \|\widetilde{\boldsymbol{X}}\Delta\|_F/\sqrt{n}\le 2\lambda_1\left(\|\Psi_0\|_{\ell_1/\ell_2}-\|\Psi_0+\Delta\|_{\ell_1/\ell_2}+c\|\Delta\|_{\ell_1/\ell_2}\right)\Big\}. \tag{3.2}$$

It is well known that when $p>n$ the matrix $\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}$ is singular and the least squares loss is not strictly convex. The design impact factor $\kappa_1(c,\Psi_0,\lambda_1,\lambda_2)$ is used to characterize the minimum curvature of the least squares loss in (2.5) when the matrix Δ is restricted in a feasible set $\mathcal{R}(c,\Psi_0,\lambda_1,\lambda_2)$. It generalizes the widely used Restricted Eigenvalue (RE) condition in high-dimensional regression [8] and is more suitable for prediction [7, 21]. We refer to Remark 3 for its connection with the RE condition.

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

$$V_\varepsilon=\mathrm{tr}(\Gamma_\varepsilon),\qquad \text{with}\quad \Gamma_\varepsilon:=\gamma_w^2\ B^{*T}\Sigma_W B^*+\gamma_e^2\ \boldsymbol{I}_m, \tag{3.3}$$

where $\mathrm{tr}(\cdot)$ stands for the trace. Let $r_e(\Gamma_\varepsilon)=\mathrm{tr}(\Gamma_\varepsilon)/\|\Gamma_\varepsilon\|_{op}$ denote the *effective rank* of Γ_ε. Write $M=n^{-1}\boldsymbol{X}^TQ_{\lambda_2}^2\boldsymbol{X}$ and $\widehat{\Sigma}=n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. Recall that P_{λ_2} and Q_{λ_2} are defined in (2.4). The following theorem provides the deviation bounds of $\|\boldsymbol{X}\widehat{F}-\boldsymbol{X}F^*\|_F$.

Theorem 4. *Under model (1.4) and Assumption 1, choose*

$$\lambda_1=4\sqrt{\max_{1\le j\le p}M_{jj}}\left(1+\sqrt{\frac{2\log(p/\epsilon')}{r_e(\Gamma_\varepsilon)}}\right)\sqrt{\frac{V_\varepsilon}{n}} \tag{3.4}$$

for any $\epsilon'>0$ and choose any $\lambda_2\ge 0$ in (2.3) such that P_{λ_2} exists. With probability $1-\epsilon-\epsilon'$,

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{F}-\boldsymbol{X}F^*\right\|_F^2\le\inf_{\substack{(\Psi_0,L_0):\\ \Psi_0+L_0=F^*}}\left[\frac{2}{n}\left\|\boldsymbol{X}(\widehat{L}-L_0)\right\|_F^2+\frac{2}{n}\left\|Q_{\lambda_2}\boldsymbol{X}(\widehat{\Psi}-\Psi_0)\right\|_F^2\right]$$

$$\le\inf_{\substack{(\Psi_0,L_0):\\ \Psi_0+L_0=F^*}}\left[4Rem_1+36\|Q_{\lambda_2}\|_{op}Rem_2(L_0)+8\|Q_{\lambda_2}\|_{op}Rem_3(\Psi_0)\right],$$

where $\|Q_{\lambda_2}\|_{op}\le 1$ and

$$Rem_1=\left(\sqrt{\mathrm{tr}(P_{\lambda_2}^2)}+\sqrt{2\log(m/\epsilon)\|P_{\lambda_2}^2\|_{op}}\right)^2\frac{V_\varepsilon}{n}$$

$$Rem_2(L_0)=\lambda_2\ \mathrm{tr}\left[L_0^T\widehat{\Sigma}(\widehat{\Sigma}+\lambda_2\boldsymbol{I}_p)^{-1}L_0\right]$$

$$Rem_3(\Psi_0)=\lambda_1^2\ \left[\kappa_1(1/2,\Psi_0,\lambda_1,\lambda_2)\right]^{-2}.$$

Since Ψ^* is not identifiable, neither Ψ^* nor L^* can be identified individually. Nevertheless, our estimator $\widehat{F}$ minimizes the error over all possible combinations of Ψ_0 and L_0 satisfying $\Psi_0 + L_0 = F^*$. As expected from (2.3), the prediction error in Theorem 4 comes from two sources: estimating L^* from the multivariate ridge regression and estimating Ψ^* from the group-lasso. Specifically, Rem_1 and $Rem_2(L_0)$ are, respectively, the variance and bias terms from the ridge regression while $Rem_3(\Psi_0)$ corresponds to the estimation error of the group-lasso. In the following, we provide more insights on these three terms.

REMARK 3 (Design impact factor and λ_1). The remainder term $Rem_3(\Psi_0)$ depends on the design impact factor $\kappa_1(c, \Psi_0, \lambda_1, \lambda_2)$ and the tuning parameter λ_1. We first discuss the connection of $\kappa_1(c, \Psi_0, \lambda_1, \lambda_2)$ and the Restricted Eigenvalue (RE) condition of $\boldsymbol{X}$ defined as

$$\kappa(s, \alpha) = \min_{S \subseteq \{1,2,\ldots,p\}, |S| \le s} \; \min_{\Delta \in \mathcal{C}(S,\alpha)} \frac{\|\boldsymbol{X}\Delta\|_F}{\sqrt{n}\|\Delta_{S\cdot}\|_F}, \tag{3.5}$$

where $\alpha \ge 1$ is a constant, $1 \le s \le p$ is some integer and $\mathcal{C}(S, \alpha) := \{\Delta \in \mathbb{R}^{p\times m} \setminus \{0\} : \alpha\|\Delta_{S\cdot}\|_{\ell_1/\ell_2} \ge \|\Delta_{S^c\cdot}\|_{\ell_1/\ell_2}\}$. Similarly, denote by $\widetilde{\kappa}(s, \alpha)$ the RE condition of $\widetilde{\boldsymbol{X}} = Q_{\lambda_2}^{1/2}\boldsymbol{X}$. In Lemma 6 of Appendix C, we show that, for any constant $c \in (0, 1)$,

$$[\kappa_1(c, \Psi_0, \lambda_1, \lambda_2)]^2 \ge \frac{[\widetilde{\kappa}(s_0, \alpha_c)]^2}{(1+c)^2 s_0} \ge \frac{\lambda_2}{\sigma_1 + \lambda_2} \cdot \frac{[\kappa(s_0, \alpha_c)]^2}{(1+c)^2 s_0}, \tag{3.6}$$

where $\alpha_c = (1 + c)/(1 - c)$, $s_0 = \|\Psi_0\|_{\ell_0/\ell_2}$ and σ_1 is the leading eigenvalue of $\widehat{\Sigma} = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$. The first inequality is proved in [21] for $m = 1$. Here we extend it to $m \ge 2$, and we further establish the second inequality which characterizes the relation between $\kappa_1(c, \Psi_0, \lambda_1, \lambda_2)$ and $\kappa(s_0, \alpha_c)$. It is well known that $\kappa(s_0, \alpha_c)$ is lower bounded by a positive constant with high probability when the rows of $\boldsymbol{X}$ are i.i.d. sub-Gaussian vectors with $\lambda_{\min}(\Sigma) > c'$ for some constant $c' > 0$ and $s_0 = O(n)$ [43]. Together with the second inequality in (3.6), we obtain $[\kappa_1(c, \Psi_0, \lambda_1, \lambda_2)]^2 \gtrsim \lambda_2/[s_0(\sigma_1 + \lambda_2)]$. Thus, when λ_2 dominates σ_1, $\kappa_1(c, \Psi_0, \lambda_1, \lambda_2)$ scales as $1/\sqrt{s_0}$.

Note that $Rem_3(\Psi_0)$ also depends on the tuning parameter λ_1 which is further related to the choice of λ_2 via the diagonal entries of $M = n^{-1}\boldsymbol{X}^T Q_{\lambda_2}^2 \boldsymbol{X}$. To simplify $Rem_3(\Psi_0)$, in Lemma 6 of Appendix C we prove

$$\max_{1\le j\le p} M_{jj} \le \max_{1\le j\le p} \widehat{\Sigma}_{jj} \left(\frac{\lambda_2}{\sigma_q + \lambda_2}\right)^2, \tag{3.7}$$

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

$$Rem_3(\Psi_0) \lesssim \frac{\lambda_2(\sigma_1+\lambda_2)}{(\sigma_q+\lambda_2)^2} \max_{1\le j\le p} \widehat{\Sigma}_{jj} \frac{s_0}{[\kappa(s_0,3)]^2}\left(1+\frac{\log(p/\epsilon')}{r_e(\Gamma_\varepsilon)}\right)\frac{V_\varepsilon}{n}.$$

The first two remainder terms Rem_1 and $Rem_2(L_0)$ depend on the choice of λ_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 $\|\boldsymbol{X}\widehat{F}-\boldsymbol{X}F^*\|_F^2$ in the following corollary. Recall that $\sigma_1 \geq \cdots \geq \sigma_q$ denote the non-zero eigenvalues of $\widehat{\Sigma}$ with $q = \mathrm{rank}(\boldsymbol{X})$.

Corollary 5. *Under model (1.4) and Assumption 1, with probability $1-\epsilon-\epsilon'$, one has*

$$\begin{aligned}\frac{1}{n}\left\|\boldsymbol{X}\widehat{F}-\boldsymbol{X}F^*\right\|_F^2 \lesssim \inf_{\substack{(\Psi_0,L_0):\\ \Psi_0+L_0=F^*}} \Bigg\{ & \frac{\sigma_1\lambda_2}{(\sigma_1+\lambda_2)(\sigma_q+\lambda_2)}\lambda_2\|L_0\|_F^2 \\ & + \left[\sum_{k=1}^q\left(\frac{\sigma_k}{\sigma_k+\lambda_2}\right)^2+\left(\frac{\sigma_1}{\sigma_1+\lambda_2}\right)^2\log(m/\epsilon)\right]\frac{V_\varepsilon}{n} \\ & + \frac{\sigma_1+\lambda_2}{\sigma_q+\lambda_2}\left(\frac{\lambda_2}{\sigma_q+\lambda_2}\right)^2 \max_{1\le j\le p}\widehat{\Sigma}_{jj}\left(1+\frac{\log(p/\epsilon')}{r_e(\Gamma_\varepsilon)}\right)\frac{s_0}{[\kappa(s_0,3)]^2}\frac{V_\varepsilon}{n}\Bigg\}\end{aligned}$$

where $\kappa(s_0,3)$ is defined in (3.5) with $s_0 = \|\Psi_0\|_{\ell_0/\ell_2}$.

Remark 4 (Orthonormal design). To draw connections with existing results on group-lasso and ridge estimators, we consider the orthonormal design $\widehat{\Sigma} = \boldsymbol{I}_p$. The deviation bounds in Corollary 5 reduce to (after ignoring the logarithmic factors)

(3.8)
$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{F}-\boldsymbol{X}F^*\right\|_F^2 \lesssim \left(\frac{1}{1+\lambda_2}\right)^2\frac{pV_\varepsilon}{n}+\left(\frac{\lambda_2}{1+\lambda_2}\right)^2\|L_0\|_F^2+\left(\frac{\lambda_2}{1+\lambda_2}\right)^2\frac{s_0V_\varepsilon}{n},$$

for any (Ψ_0, L_0) satisfying $\Psi_0+L_0 = F^*$. 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 λ_2 increases, the variance term of the ridge decreases whereas the bias term of the ridge and the error of group-lasso increase. Optimizing the right hand side of (3.8) over λ_2 yields

$$\lambda_2 = \frac{pV_\varepsilon/n}{\|L_0\|_F^2+s_0V_\varepsilon/n}. \tag{3.9}$$

(a) When $L^* = 0$, model (1.4) reduces to $Y = (\Psi^*)^T X + \varepsilon$. By choosing $(\Psi_0, L_0) = (\Psi^*, 0)$, we have $\lambda_2 = p/s_*$ from (3.9) and $\max_j M_{jj} \asymp p^2/(p + s_*)^2$ from (3.7). Consequently, the choice of λ_1 in (3.4) satisfies

$$\lambda_1 \asymp \left(\frac{p}{s_* + p}\right)^2 \left(1 + \sqrt{\frac{\log(p/\epsilon')}{r_e(\Gamma_\varepsilon)}}\right) \sqrt{\frac{V_\varepsilon}{n}}$$

and (3.8) reduces to

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

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

(b) When $\Psi^* = 0$, model (1.4) reduces to $Y = (L^*)^T X + \varepsilon$. By choosing $(\Psi_0, L_0) = (0, L^*)$, we have $\lambda_2 = pV_\varepsilon/(n\|L^*\|_F^2)$ from (3.9). After simple calculations, (3.8) yields

$$\frac{1}{n}\left\|\boldsymbol{X}\widehat{F} - \boldsymbol{X}F^*\right\|_F^2 \lesssim \min\left(\frac{pV_\varepsilon}{n}, \|L^*\|_F^2\right) \lesssim \sqrt{\frac{pV_\varepsilon}{n}}\|L^*\|_F, \tag{3.10}$$

which is the optimal rate of the ridge regression [34].

Combining scenarios (a) and (b), we conclude that the convergence rate (3.8) of our estimator $\widehat{F}$ with the optimal tuning parameters λ_1 and λ_2 matches the best possible rate even if $L^* = 0$ or $\Psi^* = 0$ were known a priori. For this reason, we refer to our estimator $\widehat{F}$ as an adaptive estimator.

When $\Psi^* = 0$ and K is much smaller than both p and m, model (1.4) is a multivariate regression with the coefficient matrix L^* exhibiting a low-rank structure. A natural approach is to use a reduced-rank estimator to estimate L^*. In Appendix D, we show that our ridge-type estimator could have a faster rate than the reduced-rank estimator in our problem.

3.2. *Statistical guarantees of estimating* Θ^*.

It is seen from (2.8) that $\widetilde{\Theta}$ depends on $\widehat{P}_{B^*}$, the estimator of P_{B^*}. In the following, we first state a general theorem which establishes the non-asymptotic upper bounds of $\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2}$ for $\widetilde{\Theta}$ obtained from (2.8) by using any estimator $\widehat{P}$ of P_{B^*} in lieu of $\widehat{P}_{B^*}$. Let Λ_1 denote the largest eigenvalue of $B^{*T}\Sigma_W B^*$.

THEOREM 6. *Under model (1.4) and Assumption 1, assume $\kappa(s_*, 4) > 0$. Let $\widetilde{\Theta}$ be any solution of problem (2.8) by using estimator $\widehat{P} \in \mathbb{R}^{m\times m}$ in place of $\widehat{P}_{B^*}$. Choose any $\lambda_3 \geq \bar{\lambda}_3$ in (2.8) with*

$$\bar{\lambda}_3 = 4\gamma_e \sqrt{\max_{1\le j\le p} \widehat{\Sigma}_{jj}}\, \frac{\sqrt{m} + \sqrt{2\log(p/\epsilon)}}{\sqrt{n}}. \tag{3.11}$$

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

$$\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lesssim \max\left\{\lambda_3,\ \frac{(\widetilde{\lambda}_3)^2}{\lambda_3}\right\} \frac{s_*}{\kappa^2(s_*, 4)}, \tag{3.12}$$

where

$$\widetilde{\lambda}_3 = \left\{\frac{1}{\sqrt{n}}\|\boldsymbol{X}F^*\|_{op} + \sqrt{\Lambda_1}\left(1 + \sqrt{\frac{K}{n}}\right)\right\} \frac{\kappa(s_*, 4)}{\sqrt{s_*}}\ \xi_n. \tag{3.13}$$

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 $\epsilon = n^{-1}$, tends to one as $n \to \infty$. The same argument is applicable to the subsequent theorems.

Theorem 6 holds for any estimator $\widehat{P}$ of P_{B^*} with convergence rate $\|\widehat{P} - P_{B^*}\|_F \lesssim \xi_n$. The effect of $\widehat{P}$ on the estimation error of $\widetilde{\Theta}$ is characterized by the term $(\widetilde{\lambda}_3)^2 s_*/[\lambda_3\kappa^2(s_*, 4)]$ in (3.12) via the choice of λ_3. When P_{B^*} can be estimated very accurately, for instance when B^* is known, ξ_n is fast enough such that $\widetilde{\lambda}_3 \leq \bar{\lambda}_3$. We can take $\lambda_3 = \bar{\lambda}_3$ to obtain the convergence rate $\bar{\lambda}_3 s_*/\kappa^2(s_*, 4)$. We refer to this as the oracle rate since it is the optimal rate for estimating Θ^* from $\boldsymbol{Y}P_{B^*}^{\perp} = \boldsymbol{X}\Theta^* + \boldsymbol{E}P_{B^*}^{\perp}$ when B^* is known (cf. [40]). On the other hand, when $\widehat{P}$ has a slow rate such that $\widetilde{\lambda}_3 > \bar{\lambda}_3$, one needs to take a larger λ_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 λ_3 is equal to $\widetilde{\lambda}_3$ and the resulting convergence rate is $\widetilde{\lambda}_3 s_*/\kappa^2(s_*, 4)$.

Remark 5 (On the benefit of group-lasso). As seen above, when $\widetilde{\lambda}_3 \leq \bar{\lambda}_3$, the convergence rate (3.12) reduces to the oracle rate $\bar{\lambda}_3 s_*/\kappa^2(s_*, 4)$. If $\log p = o(m)$ and $\max_j \widehat{\Sigma}_{jj} = O(1)$ hold, (3.11) implies $\bar{\lambda}_3 = O(\sqrt{m/n})$ by choosing $\epsilon = p^{-1}$. As a result, provided that $[\kappa(s_*, 4)]^{-1} = O(1)$, the average error per response satisfies $\sum_{j=1}^{p}[m^{-1}\sum_{\ell=1}^{m}(\widetilde{\Theta}_{j\ell} - \Theta^*_{j\ell})^2]^{1/2} = m^{-1/2}\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} = O(s_*/\sqrt{n})$, which does not depend on logarithmic factors of the feature dimension p and is faster than the standard rate of the lasso applied separately to each column of $\boldsymbol{Y}$. Such a phenomenon is known as the benefit of the group-lasso [40]. Recall that we also use the group-lasso in our first step (2.3) for estimating $\boldsymbol{X}F^*$. This benefit of the group-lasso remains and can be seen from the choice of λ_1 in (3.4). Indeed, if $\log p = o(r_e(\Gamma_\varepsilon))$ holds, by choosing $\epsilon = p^{-1}$ in (3.4), the $\log p$ term in λ_1 is negligible. The quantity $r_e(\Gamma_\varepsilon)$ is the effective rank of Γ_ε and it depends on the interplay of $B^{*T}\Sigma_W B^*$ and Σ_E. If $\lambda_1(B^{*T}\Sigma_W B^*)$ is small (e.g., upper bounded by a constant), then

$r_e(\Gamma_\varepsilon) \asymp m$, whereas if $\lambda_1(B^{*T}\Sigma_W B^*) \asymp \lambda_K(B^{*T}\Sigma_W B^*) \asymp m$, we have $r_e(\Gamma_\varepsilon) \asymp K$ and $\log p = o(r_e(\Gamma_\varepsilon))$ reduces to $\log p = o(K)$.

In the following theorem, we establish non-asymptotic upper bounds of the estimation error of our estimator $\widehat{P}_{B^*}$ obtained from Section 2.2.2. The proof is based on a variant of the Davis-Kahan theorem [48] together with careful control of the estimation error of $\widehat{\Sigma}_\varepsilon$. Let Λ_K denote the Kth largest eigenvalue of $(B^*)^T\Sigma_W B^*$.

THEOREM 7. *Under model (1.4) and Assumption 1, assume $m \le e^n$. For some constants $c, c' > 0$, one has*

$$\mathbb{P}\left\{\|\widehat{P}_{B^*} - P_{B^*}\|_F \le c \cdot Rem(P_{B^*})\right\} \ge 1 - \epsilon' - 5m^{-c'},$$

where, with V_ε and Γ_ε defined in (3.3),

$$\begin{aligned} Rem(P_{B^*}) = \inf_{\substack{(\Psi_0,L_0):\\ \Psi_0+L_0=F^*}} &\left\{ V_\varepsilon\sqrt{\frac{\log m}{n}} + \frac{\lambda_2\sigma_1}{\lambda_2+\sigma_1}\|L_0\|_F^2 + \sum_{k=1}^{q}\frac{\sigma_k}{\sigma_k+\lambda_2}\frac{V_\varepsilon}{n}\right. \\ &\left. + \frac{\lambda_2(\sigma_1+\lambda_2)}{(\sigma_q+\lambda_2)^2}\max_{1\le j\le p}\widehat{\Sigma}_{jj}\left(1+\frac{\log(p/\epsilon')}{r_e(\Gamma_\varepsilon)}\right)\frac{s_0}{[\kappa(s_0,4)]^2}\frac{V_\varepsilon}{n}\right\}. \end{aligned} \tag{3.14}$$

Recall that $\widehat{P}_{B^*}$ relies on the estimates of both Σ_ε and $\boldsymbol{X}F^*$. The first term $V_\varepsilon\sqrt{\log m/n}$ in $Rem(P_{B^*})$ is the oracle error of estimating Σ_ε in Frobenius norm even if $\boldsymbol{X}F^*$ were known. The other three terms in $Rem(P_{B^*})$ originate from the errors of estimating $\boldsymbol{X}F^*$ in Corollary 5.

When $\widetilde{\Theta}$ is obtained from (2.8) by using $\widehat{P}_{B^*}$, combining Theorem 6 and Theorem 7 yields the final rate of $\|\widetilde{\Theta}-\Theta^*\|_{\ell_1/\ell_2}$ with explicit dependency on all quantities. To simplify its expression, we introduce the parameter space

$$(\Psi^*, L^*) \in \left\{(\Psi, L) : \Psi + L = F^*, \|\Psi\|_{\ell_0/\ell_2} \le s_*, \|L\|_F^2 \le R_*\right\} \tag{3.15}$$

for some $R_* \ge 0$. Without loss of generality, we standardize the design matrix such that $\widehat{\Sigma}_{jj} = 1$ for $1 \le j \le p$. We assume the following conditions.

ASSUMPTION 2.

(a) $[\kappa(s_*, 4)]^{-1} = O(1)$, $n^{-1}\|\boldsymbol{X}F^*\|_{op}^2 = O(m + s_*)$;
(b) $\Lambda_1 \asymp \Lambda_K \asymp m$ *with* Λ_1 *and* Λ_K *being the first and* K*th eigenvalues of* $(B^*)^T\Sigma_W B^*$.

The verification of Assumption 2 is deferred to Section 3.4. Under Assumption 2, the following Corollary 8 simplifies the rates of $\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2}$

obtained by combining Theorem 6 with Theorem 7. For two sequences a_n and b_n, we write $a_n \lessapprox b_n$ for $a_n = O(b_n)$ up to multiplicative logarithmic factors of m or p. Recall that $q = \mathrm{rank}(\boldsymbol{X})$.

Corollary 8. *Under model (1.4) and Assumptions 1 & 2, assume $\kappa(s_*, 4) > 0$ and $K = O(n)$. For any (Ψ^*, L^*) satisfying (3.15), there exists a suitable choice of λ_2 in (2.3) such that the following holds with probability tending to one,*

$$\frac{1}{\sqrt{m}}\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lessapprox \max\left\{\frac{s_*}{\sqrt{n}}, \sqrt{\frac{s_*(m+s_*)}{m}} \cdot \mathrm{Err}(P_{B^*})\right\}, \tag{3.16}$$

where

$$\mathrm{Err}(P_{B^*}) = \min\left\{\frac{\sigma_1 R_*}{m} + \frac{K s_*}{n}, \frac{qK}{n}, \sqrt{\frac{(p+\sigma_1 s_*)KR_*}{nm}} + \frac{Ks_*}{n}\right\} + \frac{K}{\sqrt{n}}.$$

In view of (3.16), $s_*/\sqrt{n}$ is the oracle rate for estimating Θ^* as discussed after Theorem 6. The term $\mathrm{Err}(P_{B^*})$ quantifies the minimum price to pay for estimating P_{B^*} by $\widehat{P}_{B^*}$ over all choices of λ_2. Recalling that $\mathrm{Rem}(P_{B^*})$ in Theorem 7 depends on the choice of the tuning parameter λ_2, the derivation of $\mathrm{Err}(P_{B^*})$ minimizes $\mathrm{Rem}(P_{B^*})$ with respect to λ_2. The three error terms in $\mathrm{Err}(P_{B^*})$ correspond to different choices of λ_2 depending on the interplay of the terms in $\mathrm{Rem}(P_{B^*})$ (see the proof of Corollary 8 for more details). To facilitate understanding, we further simplify (3.16) in low- and high-dimensional settings.

Remark 6 (Further simplified rates of $\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2}$).

(i) Suppose $p < n$, $p \asymp s_*$, $\sigma_1 = O(1)$ and $K = O(\sqrt{p} \wedge \sqrt{n/p} \wedge m)$. Then (3.16) becomes

$$\begin{aligned} \frac{1}{\sqrt{m}}\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} &\lessapprox \frac{p}{\sqrt{n}}, && \text{when } p = O(m); \\ \frac{1}{\sqrt{m}}\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} &\lessapprox \frac{p}{\sqrt{n}} + \left(\frac{p}{\sqrt{n}}\right)^2, && \text{when } m = O(1). \end{aligned}$$

Recall 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 Θ^* in ℓ_1/ℓ_2 norm, our estimator $\widetilde{\Theta}$ achieves the oracle rate. If m grows, 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 $K = O(\sqrt{s_*} \wedge \sqrt{n/s_*} \wedge m)$. The upper bound in (3.16) becomes

$$\begin{aligned} &\frac{s_*}{\sqrt{n}} + \sqrt{\frac{s_* R_*}{m}} \min\left\{\sigma_1 \sqrt{\frac{R_*}{m}},\ \sqrt{\frac{(p+\sigma_1 s_*)K}{n}}\right\}, && \text{if } s_* = O(m);\\ &\frac{s_*}{\sqrt{n}} + \left(\frac{s_*}{\sqrt{n}}\right)^2 + s_*\sqrt{R_*} \min\left\{\sigma_1\sqrt{R_*},\ \sqrt{\frac{p+\sigma_1 s_*}{n}}\right\}, && \text{if } m = O(1). \end{aligned}$$

In high-dimensional case, the dimension m plays a more significant role. When m is fixed, one needs $\sigma_1 s_* R_* = o(1)$ or $s_*\sqrt{R_*(p+\sigma_1 s_*)} = o(\sqrt{n})$ for estimation consistency. This requirement is much more relaxed when $s_* = O(m)$ 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 sub-cases:

$$\begin{aligned} &\frac{1}{\sqrt{m}}\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lessapprox \frac{s_*}{\sqrt{n}} + \frac{\sqrt{s_*}\sigma_1 R_*}{m}, && \text{if } \frac{R_*}{m} \leq \frac{(p+\sigma_1 s_*)K}{n\sigma_1^2}\\ &\frac{1}{\sqrt{m}}\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2} \lessapprox \frac{s_*}{\sqrt{n}} + \sqrt{\frac{s_*(p+\sigma_1 s_*)KR_*}{nm}}, && \text{if } \frac{R_*}{m} \geq \frac{(p+\sigma_1 s_*)K}{n\sigma_1^2}. \end{aligned}$$

Intuitively, the first case is more likely to occur if σ_1, the largest eigenvalue of $\widehat{\Sigma}$, has moderate magnitude, such as $\sigma_1 = O(p/n)$. We refer to Section 3.4 for more comments on this order of σ_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 $\sigma_1 R_* = O(\sqrt{s_* n^{2\alpha-1}})$. The larger α is, the weaker the requirement on R_* becomes. On the other hand, when $\widehat{\Sigma}$ has spiked eigenvalues, for instance $\sigma_1 \asymp p$, the second case in the display above is more likely to hold. In this case, assuming $\sigma_1 \asymp p$ and $K = O(1)$, our estimator $\widetilde{\Theta}$ achieves the oracle rate $s_*/\sqrt{n}$ if $R_* = O(m/p)$. In Section 3.4, we provide examples under which $R_* = O(m/p)$ holds.

3.3. *A special case under the condition* $\Psi^* P_{B^*} = 0$. As seen in Remark 1, if $\Psi^* P_{B^*} = 0$ holds, then $\Psi^* = \Theta^*$ is identifiable. The estimator $\widehat{\Psi}$ obtained in (2.3) can be viewed as an initial estimator of Ψ^*. The convergence rate of $\|\widehat{\Psi} - \Psi^*\|_{\ell_1/\ell_2}$ is shown in Lemma 5 of Appendix B.6. Because of $\Psi^* = \Theta^*$, $\widetilde{\Theta}$ can be viewed as a refined estimator of Ψ^* and its rate of convergence has been analyzed in Theorem 6 and Corollary 8. In the following remark, we elaborate the improvement of $\widetilde{\Theta}$ over the initial estimator $\widehat{\Psi}$ in terms of their convergence rates. Empirical comparisons of these two estimators are presented in Section 6.

REMARK 7 (Comparison of $\widehat{\Psi}$ and $\widetilde{\Theta}$). Assume conditions of Corollary 5 and Assumption 2 hold. With suitable choices of λ_1 and λ_2, $\widehat{\Psi}$ obtained from (2.3) satisfies

$$\frac{1}{\sqrt{m}}\|\widehat{\Psi}-\Psi^*\|_{\ell_1/\ell_2} \lessapprox \frac{s_*\sqrt{K}}{\sqrt{n}} + \sqrt{s_*}\sqrt{\frac{\sigma_1 R_*}{m}}. \tag{3.17}$$

Comparing this rate to (3.16) (notice that $\Psi^* = \Theta^*$), the advantage of $\widetilde{\Theta}$ over the initial estimator $\widehat{\Psi}$ is substantial. For instance, in the low-dimensional case (i) of Remark 6, we have $m^{-1/2}\|\widetilde{\Theta}-\Psi^*\|_{\ell_1/\ell_2} \lessapprox p/\sqrt{n}$ provided that $p/\sqrt{n} = o(1)$. In contrast, $m^{-1/2}\|\widehat{\Psi}-\Psi^*\|_{\ell_1/\ell_2} \lessapprox p\sqrt{K/n}$ which has an extra $\sqrt{K}$ factor even if $\sigma_1 R_*/m$ is sufficiently small. In the high-dimensional case (ii), one has

$$\frac{1}{\sqrt{m}}\|\widetilde{\Theta}-\Psi^*\|_{\ell_1/\ell_2} \lessapprox \frac{s_*}{\sqrt{n}} + \sqrt{\frac{s_*(m+s_*)}{m}}\left(\frac{\sigma_1 R_*}{m} + \sqrt{\frac{s_*}{n}}\right)$$

which is always faster than (3.17) provided that $\sigma_1 R_* = o(m/s_*)$ and $s_* = O(mK)$. Note that in (3.17) we need $\sigma_1 R_* = o(m/s_*)$ for the consistency of $\widehat{\Psi}$. For further illustration, suppose $s_* = O(m)$, $m \asymp n^\alpha$ for some $\alpha \geq 1/2$ and $\sigma_1 R_* = O(\sqrt{s_* n^{2\alpha-1}})$, then $m^{-1/2}\|\widetilde{\Theta}-\Psi^*\|_{\ell_1/\ell_2} \lessapprox s_*/\sqrt{n}$ corresponds to the oracle rate, whereas (3.17) becomes $m^{-1/2}\|\widehat{\Psi}-\Psi^*\|_{\ell_1/\ell_2} \lessapprox s_*(K/n)^{1/2} + \sqrt{s_*}(s_*/n)^{1/4}$.

3.4. *Validity of Assumption 2 and conditions in Remark 6.*

In this section we provide theoretical justifications for Assumption 2 as well as some conditions on σ_1 and the radius R_* that we mentioned in Remark 6.

Part (a) of Assumption 2 contains standard conditions on the design matrix. Suppose the rows of $\boldsymbol{X}\Sigma^{-1/2}$ are i.i.d. sub-Gaussian random vectors with bounded sub-Gaussian constant. The validity of $[\kappa(s_*, 4)^{-1} = O(1)$ is already discussed in Remark 3. Regarding condition $n^{-1}\|\boldsymbol{X}F^*\|_{op}^2 = O(m + s_*)$, suppose $s_* = O(n)$, $K = O(n)$ and $\lambda_1(\Sigma_Z) = O(1)$. We show in Lemma 8 of Appendix C that $n^{-1}\|\boldsymbol{X}F^*\|_{op}^2 = O_p(m + s_*)$ provided that $\|\Psi^*\|_{op}^2 = O(s_* + m)$, $\|B^*\|_{op}^2 = O(m)$ and $\|\Sigma_{S_*S_*}\|_{op} = O(1)$ where S_* is the set of nonzero rows of Ψ^*. Recall that $\|\Psi^*\|_{\ell_0/\ell_2} \leq s_*$ and $B^* \in \mathbb{R}^{K\times m}$ with $K < m$, $\|\Psi^*\|_{op}^2 = O(m + s_*)$ and $\|B^*\|_{op}^2 = O(m)$ hold when either $\|\Psi^*\|_\infty = O(1)$ and $\|B^*\|_\infty = O(1)$ or entries of Ψ^* and B^* are i.i.d. samples from a mean-zero distribution with bounded fourth moment [6].

Condition (b) is standard when m (and also K) is fixed. When m grows with n, we note that $\boldsymbol{\varepsilon} = \boldsymbol{W}B^* + \boldsymbol{E}$ follows a factor model where $\boldsymbol{W}$ 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^* [4, 25, 26, 27]. In particular, condition (b) holds if $c \le \lambda_K(\Sigma_W) \le \lambda_1(\Sigma_W) \le C$ 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 $\widetilde{\lambda}_3$ in (3.13) and $Rem(P_{B^*})$ in (3.14). If Λ_1 and Λ_K have different rates, we can replace them by the corresponding rates and simplify the error bounds of $\widetilde{\Theta}$ accordingly. We also verify the empirical performance of our procedure in Appendix E when some of $\Lambda_1, \ldots, \Lambda_K$ are moderate or small.

Regarding conditions on σ_1 in part (ii) of Remark 6, when $\|\Sigma\|_{op} = O(1)$, one has $\sigma_1 = O_p(p/n)$ by $\|\widehat{\Sigma} - \Sigma\|_{op} = O_p(\sqrt{p/n} \vee (p/n))$ from [46]. To see when $\sigma_1 \asymp p$ holds, suppose $\|\Sigma\|_{op} \asymp p$ and $\log p = o(n)$. Since $\|\widehat{\Sigma} - \Sigma\|_{op} \le \|\widehat{\Sigma} - \Sigma\|_F = O_p(p\sqrt{\log p/n})$ (for instance, see the argument in Lemma 14 of Appendix C.6), one can deduce that $\sigma_1 \asymp p$ with high probability.

In the end, we comment on the magnitude of R_*, the upper bound of $\|L^*\|_F^2$ in (3.15). In the mediation analysis via structural equation models, $L^* = A^*B^*$ is known as the indirect effect of X on Y. Under model (1.4), the estimation of the non-sparse coefficient matrix $\Psi^* + L^*$ becomes challenging in high dimension when $\|L^*\|_F^2$ is large, and the estimation error is further accumulated in the rates of the final estimator $\widetilde{\Theta}$ as shown in Corollary 8. Thus, intuitively $\|L^*\|_F^2$ cannot grow too fast in order to guarantee the consistency of $\widetilde{\Theta}$. This can be compared to the standard results in linear regression. For instance, in linear regression $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ where $\boldsymbol{y} \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ is dense, one needs $\|\beta\|_2^2 = o(1)$ for consistent estimation when $p > n$; see [34, 23] for the minimax lower bound.

In the following, we discuss under what conditions $\|L^*\|_F^2$ is small and how small it can be. Provided that $\|B^*\|_{op}^2 = O(m)$, we first have $\|L^*\|_F^2/m \le \|A^*\|_F^2\|B^*\|_{op}^2/m = O(\|A^*\|_F^2)$. Since $A^* = \Sigma^{-1}\mathrm{Cov}(X, Z)$, intuitively $\|A^*\|_F$ is small when either (1) $\mathrm{Cov}(X, Z)$ is close to zero or (2) Σ is "large".

To show when case (1) holds, suppose the smallest eigenvalue of Σ is bounded away from zero. When $\mathrm{Cov}(X, Z)$ is sparse with $\|\mathrm{Cov}(X, Z)\|_{\ell_0} = O(1)$ and $\max_{j,k} |\mathrm{Cov}(X_j, Z_k)| \lesssim \xi$, one has $\|A^*\|_F^2 = O(\xi^2)$ which vanishes if $\xi = o(1)$. When $\mathrm{Cov}(X, Z)$ is dense with $\|\mathrm{Cov}(X, Z)\|_{\ell_0} \ge c'(pK)$ for some small constant $c' > 0$, if the range of the nonzero entries of $\mathrm{Cov}(X, Z)$ is bounded, an application of Pólya-Szegö's inequality (see, for instance, [24]) yields $\|\mathrm{Cov}(X, Z)\|_F \lesssim \|\mathrm{Cov}(X, Z)\|_{\ell_1/\ell_1}/\sqrt{pK}$. Therefore, provided that $\|\mathrm{Cov}(X, Z)\|_{\ell_1/\ell_1} = O(1)$, one has $\|L^*\|_F^2/m = O(\|A^*\|_F^2) = O(1/(pK))$.

To show when case (2) holds, we consider the setting that X follows an approximate factor model $X = \Gamma F + W'$, where the noise W' and the factor F are independent, $\mathrm{Cov}(F)$ and $\mathrm{Cov}(W')$ have bounded eigenvalues and the loading matrix $\Gamma \in \mathbb{R}^{p\times\bar{K}}$ satisfies the pervasiveness assumption $\lambda_{\bar{K}}(\Gamma\Gamma^T) \gtrsim p$. The number of factors, $\bar{K}$, is often much smaller than p. In this scenario, $\Sigma = \Gamma\mathrm{Cov}(F)\Gamma^T + \mathrm{Cov}(W')$ has $\bar{K}$ 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 $V_{(\bar{K})} = (v_1, \ldots, v_{\bar{K}}) \in \mathbb{R}^{p\times\bar{K}}$ and $V_{(-\bar{K})} = (v_{\bar{K}+1}, \ldots, v_p) \in \mathbb{R}^{p\times(p-\bar{K})}$. Provided that

$$\left\|V_{(\bar{K})}^T\mathrm{Cov}(X,Z)\right\|_{op} \leq c\sqrt{p} \quad \text{and} \quad \left\|V_{(-\bar{K})}^T\mathrm{Cov}(X,Z)\right\|_{op} \leq c/\sqrt{p}, \tag{3.18}$$

for some sufficiently small constant $c > 0$, we obtain

$$\begin{aligned}\|A^*\|_{op} &= \left\|\Sigma^{-1}\mathrm{Cov}(X,Z)\right\|_{op} \\ &\leq \frac{1}{d_{\bar{K}}}\left\|V_{(\bar{K})}^T\mathrm{Cov}(X,Z)\right\|_{op} + \frac{1}{d_p}\left\|V_{(-\bar{K})}^T\mathrm{Cov}(X,Z)\right\|_{op} = O(1/\sqrt{p}).\end{aligned}$$

We thus have$\|L^*\|_F^2/m = O(\|A^*\|_F^2) = O(K/p)$. Also notice that this setting does not conflict with part (b) of Assumption 2. Indeed, provided that $c \leq \lambda_K(\Sigma_Z) \leq \lambda_1(\Sigma_Z) \leq C$ and $cm \leq \lambda_K(B^*B^{*T}) \leq \lambda_1(B^*B^{*T}) \leq Cm$ for some constants $c, C > 0$, one can deduce $c/2 \leq \lambda_K(\Sigma_W) \leq \lambda_1(\Sigma_W) \leq C$ from $\Sigma_W = \Sigma_Z - \mathrm{Cov}(Z,X)\Sigma^{-1}\mathrm{Cov}(X,Z)$.

Condition (3.18) requires that: (1) the order of $\|\mathrm{Cov}(X,Z)\|_{op}$ cannot be greater than $\sqrt{p}$; (2) the columns of $\mathrm{Cov}(X,Z)$ and $V_{(-\bar{K})}$ are approximately orthogonal. From a practical perspective, under the structural equation model (1.2), condition (3.18) implies that the causal effect of X on Z (i.e., the matrix A^*) is weak due to the spiked eigenvalues of Σ in high dimension. However, in view of the factor model $X = \Gamma F + W'$, the association between the hidden variable Z and the low dimensional factor F can be strong. Though $\|L^*\|_F$ or $\|A^*\|_{op}$ is required to be small when p is large, our analysis allows nontrivial dependence between Z and F, which could be reasonable in many practical situations.

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, $\Sigma_E = \mathrm{diag}(\tau_1^2, \ldots, \tau_m^2)$, and discuss how to modify our approach correspondingly.

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

$$\Sigma_\varepsilon = (B^*)^T \Sigma_W B^* + \mathrm{diag}(\tau_1^2, \ldots, \tau_m^2). \tag{4.1}$$

In contrast to the homoscedastic case, the eigenspace of Σ_ε corresponding to the first K eigenvalues, in general, no longer coincides with the row space of B^*. Consequently, one cannot identify P_{B^*} via the eigenspace of Σ_ε as in Section 2.1. To overcome this difficulty, we resort to a newly developed procedure called HeteroPCA proposed by [50]. For completeness, we restate their procedure in Algorithm 2. The main idea is to iteratively perform the singular value decomposition (SVD) on the estimates of Σ_ε to impute its diagonal. Under a mild incoherence condition on the row space of B^*, P_{B^*} can be recovered by applying Algorithm 2 to Σ_ε. Thus Θ^* is identifiable from (2.2). We summarize the identifiability below in Proposition 9.

Recall that $P_{B^*} = UU^T$ with $U := U(K) \in \mathbb{R}^{m \times K}$ being the first K right singular vectors of B^*, and Λ_1 and Λ_K are the first and Kth eigenvalues of $(B^*)^T \Sigma_W B^*$, respectively. Let $\{e_j\}_{j=1}^m$ denote the canonical basis of $\mathbb{R}^m$.

PROPOSITION 9. *Under model (1.4), assume $\Sigma_E = diag(\tau_1^2, \ldots, \tau_m^2)$ and rank$(\Sigma_W) = K$. Further assume*

$$\frac{\Lambda_1}{\Lambda_K} \max_{1 \le j \le m} \|e_j^T U\|_2^2 \le C_U \tag{4.2}$$

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

An application of Theorem 3 in [50] guarantees the recovery of P_{B^*} from Σ_ε 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^*}, 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 $(B^*)^T \Sigma_W B^*$ from a diagonal matrix with unequal entries. We also note that $\frac{m}{K} \max_{1 \le j \le m} \|e_j^T U\|_2^2$ is known as the *incoherence constant* in the matrix completion literature [16, 17]. When $\Lambda_1 \asymp \Lambda_K$, (4.2) requires $\max_{1 \le j \le m} \|e_j^T U\|_2^2 = O(1)$ which is much weaker than the typical incoherence condition $\max_{1 \le j \le m} \|e_j^T U\|_2^2 = O(K/m)$, assumed in the matrix completion literature. Finally, Proposition 3 in [50] implies that condition (4.2) in general cannot be further relaxed in order to recover P_{B^*} from Σ_ε.

Remark 8 (Identification via PCA when $m \to \infty$). We propose to use HeteroPCA to identify P_{B^*} in the presence of heteroscedasticity since it guarantees the identifiability of Θ^* for any $m > K$ under condition (4.2). Directly applying PCA to Σ_ε as in Section 2.1 may not recover P_{B^*} hence not identify Θ^*. However, we remark that PCA is robust against the departure from homoscedasticity, and even from the diagonal structure of Σ_E, when Λ_K, the Kth eigenvalue of $(B^*)^T \Sigma_W B^*$, diverges fast enough as $m \to \infty$. Specifically, at the population level, applying PCA to Σ_ε identifies P_{B^*} asymptotically provided that $\sqrt{K}\|\Sigma_E\|_{op} = o(\Lambda_K)$, as $m \to \infty$. This phenomenon is known as the blessing of dimensionality in the factor model literature [4, 26, 27]. Most of the SVA methods, for instance [37, 41], rely on this robustness of PCA. Their methods thus only guarantee the asymptotic identifiability when $m \to \infty$, and are not applicable if m is fixed.

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

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

4.2. *Estimation.* Our estimation procedure under heteroscedasticity remains the same except estimating U by HeteroPCA in Algorithm 2. To be specific, we consider the estimator $\widetilde{P}_{B^*} = \widetilde{U}\widetilde{U}^T$, where $\widetilde{U}$ is obtained from Algorithm 2 with the input $\widehat{\Sigma} = \widehat{\Sigma}_\varepsilon$, $r = K$ 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).

4.3. *Statistical guarantees.* Our estimation algorithm enjoys similar statistical guarantees as in Section 3. First, since $\widehat{F}$ is the same estimator obtained from (2.3), the deviation bounds of $\|\boldsymbol{X}\widehat{F} - \boldsymbol{X}F^*\|_F$ in Theorem 4 and Corollary 5 still hold under Assumption 1. Second, Theorem 10 below

Algorithm 3 The H-HIVE procedure for estimating Θ^*.

```
Require: Data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times m}$, rank $K$, number of iterations $T$, tuning
    parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$.
1: Estimate $\boldsymbol{X}\widehat{F}$ with $\widehat{F} = \widehat{\Psi} + \widehat{L}$ by solving (2.3).
2: Obtain $\widehat{\Sigma}_\varepsilon$ from (2.7).
3: Compute $\widetilde{P}_{B^*} = \widetilde{U}\widetilde{U}^T$ with $\widetilde{U}$ obtained from HeteroPCA($\widehat{\Sigma}_\varepsilon, K, T$) in Algorithm 2.
4: Estimate $\Theta^*$ by solving (2.8) with $\widetilde{P}_{B^*}$ in lieu of $\widehat{P}_{B^*}$.
```

provides non-asymptotic upper bounds for $\|\widetilde{P}_{B^*} - P_{B^*}\|_F$ with $\widetilde{P}_{B^*} = \widetilde{U}\widetilde{U}^T$ and $\widetilde{U}$ obtained from Algorithm 2. Finally, since $\widetilde{\Theta}$ is obtained from the same criterion in (2.8) by using $\widetilde{P}_{B^*}$ in place of $\widehat{P}_{B^*}$, the convergence rate of $\|\widetilde{\Theta} - \Theta^*\|_{\ell_1/\ell_2}$ immediately follows from the following theorem in conjunction with Theorem 6.

Theorem 10. *Under the same conditions of Theorem 7, assume condition (4.2) holds and $Rem(P_{B^*}) \le c\sqrt{K}$ for some constant $c > 0$ with $Rem(P_{B^*})$ defined in (3.14). For some constants $c', c'' > 0$, the estimator $\widetilde{P}_{B^*} = \widetilde{U}\widetilde{U}^T$ with $\widetilde{U}$ obtained from Algorithm 2 satisfies*

$$\mathbb{P}\left\{\|\widetilde{P}_{B^*} - P_{B^*}\|_F \le c' \cdot Rem(P_{B^*})\right\} \ge 1 - \epsilon' - 5m^{-c''}.$$

The proof of Theorem 10 mainly relies on a new robust $\sin\Theta$ theorem stated in Appendix A, which provides upper bounds for the Frobenius norm of $\sin\Theta(\widetilde{U}, U) := \widetilde{U}_\perp^T U$, where $\widetilde{U}$ is the output of Algorithm 2 and $\widetilde{U}_\perp$ is its orthogonal complement. The new $\sin\Theta$ theorem complements Theorem 3 in [50] which controls the operator norm of $\sin\Theta(\widetilde{U}, U)$. In order to establish the rate of $\widetilde{\Theta}$, we need this new result to control the Frobenius norm of the estimated eigenspace. This technical tool can be of its own interest and is potentially useful for many other problems.

The validity of Theorem 10 also hinges on the condition $Rem(P_{B^*}) \le c\sqrt{K}$. Under conditions of Corollary 8 and Remark 6, by inspecting their proofs in Appendix C.3, one can verify that $Rem(P_{B^*}) = O(\sqrt{K})$ holds for a suitable choice of λ_2 provided that, up to a multiplicative logarithmic factor, $p\sqrt{K} = O(n)$ in the low-dimensional case or $(s_* \vee \sqrt{K})\sqrt{K} = O(n)$ and $\sigma_1 R_* = O(m\sqrt{K})$ in the high-dimensional case.

Remark 9 (Effect of heteroscedasticity on estimating Θ). Heteroscedasticity affects the estimation error of $\widetilde{\Theta}$ implicitly via V_ε defined in (3.3) and $\bar{\lambda}_3$ in (3.11). For simplicity of presentation, we assumed $\{\boldsymbol{E}_{ij}\}_{j=1}^m$ shares the same sub-Gaussian constant γ_e in Assumption 1. To illustrate the effect of heteroscedasticity, one could instead assume $\boldsymbol{E}_{ij}/\tau_j$ is γ_e sub-Gaussian for

$1 \le j \le m$. Then by inspecting the proof and using modified arguments in Lemmas 9 – 12 in Appendix C.6, it is straightforward to show that the same results in Theorems 4, 6, 7 and 10 hold with V_ε and $\bar{\lambda}_3$ replaced by

$$V'_\varepsilon = \gamma_w^2 \mathrm{tr}\left(B^{*T}\Sigma_W B^*\right) + \gamma_e^2 m\bar{\tau}^2, \quad \bar{\lambda}'_3 = 4\gamma_e\bar{\tau}\sqrt{\max_{1\le\ell\le p}\widehat{\Sigma}_{\ell\ell}}\frac{\sqrt{m}+\sqrt{2\log(p/\epsilon)}}{\sqrt{n}}.$$

where $\bar{\tau}^2 = m^{-1}\sum_{j=1}^m \tau_j^2$. The quantity $\bar{\tau}^2$ reduces to τ^2 in the homoscedastic case. But in the presence of strong heteroscedasticity, $\bar{\tau}^2$ can be of order different from $O(1)$.

To conclude this section, we compare the estimation errors of $\widetilde{P}_{B^*}$ and the PCA-based estimator $\widehat{P}_{B^*}$ in Section 2.2.2 in the presence of heteroscedasticity. Recall that Λ_K denotes the Kth eigenvalue of $B^{*T}\Sigma_W B^*$.

Theorem 11. *Suppose the same conditions of Theorem 7 hold. Then*

$$\mathbb{P}\left\{\|\widehat{P}_{B^*} - P_{B^*}\|_F \le c \cdot Rem^{(h)}(P_{B^*})\right\} \ge 1 - \epsilon' - 5m^{-c'}$$

for some constants $c, c' > 0$, where

$$Rem^{(h)}(P_{B^*}) = Rem(P_{B^*}) + \frac{1}{\Lambda_K}\left[\sum_{j=1}^m \left(\tau_j^2 - \bar{\tau}^2\right)^2\right]^{1/2} \tag{4.3}$$

with $Rem(P_{B^*})$ defined in (3.14).

Comparing (4.3) with (3.14), the last term in (4.3) is the bias of PCA due to heteroscedasticity and it is exactly zero when the error E is homoscedastic. In general, this bias term could vanish if Λ_K is large and the degree of heteroscedasticity is small, such as $\Lambda_K \gtrsim m$ and $\sum_{j=1}^m(\tau_j^2 - \bar{\tau}^2)^2 = O(m)$ as $m \to \infty$. This can be viewed as the sample analog of the robustness of PCA in Remark 8. However, we note that, even if the bias term converges to 0, it may have a slower rate than $Rem(P_{B^*})$ which renders the rate of $\widehat{P}_{B^*}$ slower than that of $\widetilde{P}_{B^*}$ from HeteroPCA.

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 Θ^*. Selection of tuning parameters and recommendation of standardization are discussed subsequently. In the end, we discuss in details the practical usage of our estimator of Θ^* for inferring Ψ^*.

5.1. *Selection of K.* Recall that $\boldsymbol{\varepsilon} = \boldsymbol{W}B^* + \boldsymbol{E}$ and K corresponds to the rank of the unknown coefficient matrix B^*. When $\boldsymbol{\varepsilon}$ and $\boldsymbol{W}$ are both observable, estimation of the rank of coefficient matrix has been studied by [13, 14, 29, 9] in the framework of multivariate regression. However, since $\boldsymbol{\varepsilon}$ and $\boldsymbol{W}$ are both unobserved, we view $\boldsymbol{\varepsilon} = \boldsymbol{W}B^* + \boldsymbol{E}$ as a factor model with K being the number of factors. [5] 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 Σ_E in view of (4.1), the theoretical justification of this class of methods is unclear under our model.

In the following we consider an eigenvalue ratio approach originally developed by [36, 1] for factor models. Specifically, we estimate $\boldsymbol{\varepsilon}$ by $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{F}$ with $\widehat{F}$ obtained from (2.3) and construct $\widehat{\Sigma}_\varepsilon$ as (2.7). We then propose to estimate K by

$$\widehat{K} = \arg\max_{j\in\{1,2,\ldots,\bar{K}\}} \widehat{\lambda}_j/\widehat{\lambda}_{j+1}, \tag{5.1}$$

where $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots$ are the eigenvalues of $\widehat{\Sigma}_\varepsilon$ and $\bar{K}$ is a pre-specified number, for example, $\bar{K} = \lfloor (n \wedge m)/2 \rfloor$ [36] with $\lfloor x \rfloor$ standing for the largest integer that is no greater than x. This procedure does not require the knowledge of any unknown quantity, such as the noise level $\|\Sigma_E\|_{op}$. The following theorem provides theoretical justification for the above procedure.

THEOREM 12. *Under model (1.1) or equivalently (1.4) with heteroscedastic noise $\Sigma_E = diag(\tau_1^2, \ldots, \tau_m^2)$, suppose condition (b) in Assumption 2 holds. Assume $\max_{1\leq j\leq m}\tau_j^2 = O(1)$, $Rem(P_{B^*}) = o(1)$ with $Rem(P_{B^*})$ defined in (3.14). Then with probability $1 - \epsilon' - 5m^{-c''}$ for some constant $c'' > 0$,*

$$\frac{\widehat{\lambda}_j}{\widehat{\lambda}_{j+1}} \asymp 1, \quad \textit{for } 1 \leq j \leq K-1, \ \textit{ and } \quad \frac{\widehat{\lambda}_{K+1}}{\widehat{\lambda}_K} = O\left(Rem(P_{B^*}) + m^{-1}\right).$$

Under Assumption 2, $K = O(1)$, $s_* = o(n)$ and $\sigma_1 R_* = o(m)$, one can deduce from the proof of Corollary 8 that $Rem(P_{B^*}) = o(1)$ for a suitable choice of λ_2. In addition, if $m \to \infty$, we obtain $\widehat{\lambda}_K/\widehat{\lambda}_{K+1} \to \infty$. Thus, the maximizer of $\widehat{\lambda}_j/\widehat{\lambda}_{j+1}$ is no smaller than K asymptotically, i.e., $\widehat{K} \geq K$, 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 in PCA. If we plot the ratio $\widehat{\lambda}_j/\widehat{\lambda}_{j+1}$ against j, we expect that the curve has a sharp increase at $j = K$ (since $\widehat{\lambda}_K/\widehat{\lambda}_{K+1} \to \infty$), 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 yield the same results. In Section 6.2, we conduct extensive simulations to compare our criterion (5.1) with some other existing methods for selecting K [12].

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 Θ^*. Recall that, after estimating K by $\widehat{K}$, we construct $\widehat{P}_{\widehat{K}} = \widehat{U}_{\widehat{K}}\widehat{U}_{\widehat{K}}^T$ and use it in lieu of $\widehat{P}_{B^*}$ in (2.8) to estimate Θ^*. For illustration purpose, we consider the case that $\widehat{K} = r$ for some fixed integer $1 \le r \le m$. At the population level, suppose 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 $r \ge K$, the first K columns of U_r align with those of U_K and the rest of $r-K$ columns are arbitrary but orthogonal to U_K. Similar to $P_{B^*} = U_K U_K^T$, $P_r = U_r U_r^T$ is also a projection matrix. The following lemma demonstrates that the effect of using P_r to estimate Θ^* is characterized by the difference of two projection matrices P_r and P_{B^*}.

LEMMA 2. *Under model (1.4), $P_r^{\perp} Y$ is equal to*

$$\begin{cases} \left[\Theta^* + \left(\Psi^* P_{B^*} + A^* B^*\right)\left(P_{B^*} - P_r\right)\right]^T X + P_r^{\perp}[(B^*)^T W + E], & \textit{if } r < K; \\ (\Theta^*)^T X - (P_r - P_{B^*})(\Theta^*)^T X + P_r^{\perp} E, & \textit{if } r > K; \\ (\Theta^*)^T X + P_r^{\perp} E, & \textit{if } r = K. \end{cases}$$

As we can see, if $r < K$, the estimand of (2.8) is $\Theta^* + (\Psi^* P_{B^*} + A^* B^*)(P_{B^*} - P_r) = \Theta^* + [\Psi^* B^{*T}(B^* B^{*T})^{-1} + A^*](B^*)_{(-r)}$ 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 $[\Psi^* B^{*T}(B^* B^{*T})^{-1} + A^*](B^*)_{(-r)}$. Intuitively, if the last $K - r$ singular values of B^*, $d_{r+1}, \ldots, d_K$, are relatively small and close to zero, we expect the bias to be negligible. In this case, underestimating K may still lead to a reasonably accurate estimate of Θ^*. On the other hand, if $r > K$, our estimator is also biased, and the bias is equal to $-(P_r - P_{B^*})(\Theta^*)^T = -P_r(\Theta^*)^T$ (the equality holds by the orthogonality between P_{B^*} and $P_{B^*}^{\perp}$). Its magnitude depends on the angle between rows of Θ^* and the last $r - K$ columns of U_r.

5.3. *Choosing tuning parameters λ_1, λ_2 and λ_3.* Recall that our procedure (Algorithms 1 and 3) requires three tuning parameters λ_1, λ_2 and λ_3. Since the first two parameters (λ_1, λ_2) and the third one λ_3 appear in two optimization problems (2.3) and (2.8), respectively, we propose to select (λ_1, λ_2) and λ_3 separately by cross validation. When estimating F^* in (2.3), we can search λ_1 and λ_2 over a two-way grid to minimize the mean squared prediction error via k-fold cross validation.[3] Similarly, when estimating Θ^* in (2.8), we can tune λ_3 by k-fold cross validation over a grid of λ_3.

5.4. *Standardization.* In steps (2.3) and (2.8) of our estimation procedure, the tuning parameters λ_1 and λ_3 depend on $\max_{1\le j\le p} \widehat{\Sigma}_{jj}$ from Theorems 4 and 6. This dependency comes from the union bounds argument for controlling $\max_{1\le j\le p} \|\boldsymbol{X}_j^T P_{\lambda_2}\boldsymbol{\varepsilon}\|_2$. To tighten the bound in practice, we recommend standardizing the columns of $\boldsymbol{X}$ to unit variance. Since the means of $\boldsymbol{Y}$ and $\boldsymbol{X}$ do not affect the estimation of Θ^*, one can center both $\boldsymbol{X}$ and $\boldsymbol{Y}$ before fitting the model.

5.5. *Practical usage of Θ^* for inferring Ψ^*.* When the parameter Ψ^* is of primary interest, the information in the parameter $\Theta^* = \Psi^* P_{B^*}^{\perp}$ is still helpful to infer Ψ^*. In the following, we discuss this usage of Θ^* in two scenarios. The first scenario corresponds to $\Theta^* \approx \Psi^*$ whence one can use $\widehat{\Theta}$ to estimate Ψ^*. We also provide sufficient conditions for $\Theta^* \approx \Psi^*$. We then suggest further usage of Θ^* when $\Theta^* \not\approx \Psi^*$ in the second scenario. Finally, we offer our recommendations to practitioners.

Case (1): $\Theta^* \approx \Psi^*$. In this case, our estimator of Θ^* also estimates Ψ^* consistently. To see when $\Theta^* \approx \Psi^*$ holds, recall that $\Psi^* - \Theta^* = \Psi^* P_{B^*}$. Then $\Psi^* \approx \Theta^*$ is implied by $\Psi^* P_{B^*} \to 0$ as $m \to \infty$. The following two lemmas provide different sets of sufficient conditions for $\max_{1\le j\le p} \|P_{B^*}\Psi^*_{j\cdot}\|_2 = o(\sqrt{m})$ whence

$$\frac{1}{\sqrt{m}} \max_{1\le j\le p} \|\Psi^*_{j\cdot} - \Theta^*_{j\cdot}\|_2 = o(1).$$

Their proofs can be found in Appendix C.5. Recall that $U \in \mathbb{R}^{m\times K}$ contains the right singular vectors of $B^* \in \mathbb{R}^{K\times m}$.

LEMMA 3. *Suppose the columns of U are uniformly distributed over the families of K orthonormal vectors. Provided that $K = o(m)$, for any $1 \le j \le p$, one has*

$$\left\|P_{B^*}\Psi^*_{j\cdot}\right\|_2^2 = O_p\left(\left\|\Psi^*_{j\cdot}\right\|_2^2 \frac{K}{m}\right).$$

[3] When both p and m are large, searching (λ_1, λ_2) over a fine two-way grid could be computationally intensive, we offer an alternative way of selecting λ_1 and λ_2 in Appendix E which costs less computation.

If $\|\Psi^*\|_\infty = O(1)$ holds additionally, then

$$\frac{1}{m}\left\|P_{B^*}\Psi^*_{j\cdot}\right\|_2^2 = O_p\left(\frac{K}{m}\right).$$

Lemma 3 states that when the directions of columns of U are random enough (more specifically, uniformly distributed) and $K = o(m)$, the matrix P_{B^*} is incoherent to $\Psi^*_{j\cdot}$. The uniformity assumption of U is commonly made in the matrix completion literature, under which [17, 16] prove that $\max_{1\le j\le m}\|P_{B^*}e_j\|_2^2 = O_p(K/m)$ where $\{e_j\}_{1\le j\le m}$ is the canonical basis of $\mathbb{R}^m$. Our result in Lemma 3 reduces to this existing result when $\Psi^*_{j\cdot}$ is aligned with canonical vectors.

LEMMA 4. *Suppose $\max_{1\le j\le p}\|\Psi^*_{j\cdot}\|_0 \le d$ and $\|\Psi^*\|_\infty \le c$ for some constant $c > 0$ and integer $1 \le d \le m$. Further assume $\lambda_K(m^{-1}B^*B^{*T}) \ge c'$ and $\|B^*\|_\infty \le c''$ for some constants $c', c'' > 0$. Then, for any $1 \le j \le p$,*

$$\frac{1}{m}\left\|P_{B^*}\Psi^*_{j\cdot}\right\|_2^2 \le \frac{(cc'')^2}{c'}\frac{dK}{m^2}.$$

From Lemma 4, under certain regularity conditions on Ψ^* and B^*, one has $m^{-1}\|P_{B^*}\Psi^*_{j\cdot}\|_2^2 = o(1)$ if $dK/m^2 = o(1)$ which holds when either $\Psi^*_{j\cdot}$ is sufficiently sparse or K is much smaller than m. Sparsity of rows of Ψ^* is commonly assumed in the SVA literature and is practically meaningful in many biological applications, see, for instance, [28, 47, 41]. In particular, similar sets of sufficient conditions for $m^{-1}\|P_{B^*}\Psi^*_{j\cdot}\|_2^2 = o(1)$ are given in [41, 47].

Case (2): $\Theta^* \not\approx \Psi^*$. We provide two ways of using the estimator of Θ^* to infer Ψ^* in this scenario.

(i) Suppose we are interested in Ψ^*C for some known constraint matrix $C \in \mathbb{R}^{m\times q}$. Then provided that $P_{B^*}C$ is small, one could use $\widehat{\Theta}C$ to estimate Ψ^*C because $\Theta^*C = \Psi^*C - \Psi^*P_{B^*}C \approx \Psi^*C$. In practice, since P_{B^*} can be estimated by $\widehat{P}_{B^*}$ or $\widetilde{P}_{B^*}$ (see Section 2.2.2 for homoscedastic error and Section 4.2 for heteroscedastic error), researchers could empirically decide whether or not using $\widehat{\Theta}C$ to estimate Ψ^*C by comparing the magnitude of $\|\widehat{P}_{B^*}C\|_F/\sqrt{mq}$ with a small tolerance level. As a simple yet important example, suppose $B^* = [B_1\ \mathbf{0}]$ for some $B_1 \in \mathbb{R}^{K\times m_1}$ and $1 \le m_1 \le m$. From model (1.1), this structure of B^* implies that there are m_1 responses affected by the hidden variables Z. Let $S \subseteq \{1,\dots,m\}$ denote the index set of columns of B_1 and write

$S^c = \{1, \dots, m\} \setminus S$. Since the set S can be estimated from the sparsity pattern of $\widehat{P}_{B^*}$, we assume S is known for simplicity. Then it is easy to see that, for any $1 \le j \le p$,

$$\Theta^*_{j\cdot} = P^{\perp}_{B^*} \Psi^*_{j\cdot} = \begin{bmatrix} P^{\perp}_{B_1} \Psi^*_{jS} \\ \Psi^*_{jS^c} \end{bmatrix},$$

which further implies $\Theta^*_{j\ell} = \Psi^*_{j\ell}$ for any $\ell \in S^c$ and $1 \le j \le p$. Intuitively, since the ℓth response is not associated with hidden variables, the parameter $\Psi^*_{\cdot\ell} \in \mathbb{R}^p$ in the multivariate regression is identifiable and is indeed identical to our estimand $\Theta^*_{\cdot\ell}$. In this case, for any given constraint matrix $C = [\mathbf{0}\ C_1]^T \in \mathbb{R}^{m \times q}$ with $C_1 \in \mathbb{R}^{q \times (m - m_1)}$ and the index set of rows of C_1^T in C being S^c, we can use our estimator $\widehat{\Theta} C$ to infer $\Psi^* C$.

(ii) Another usage of Θ^* is to further infer the non-zero rows of Ψ^* based on the fact that $\Theta^*_{j\cdot} \ne 0$ implies $\Psi^*_{j\cdot} \ne 0$. Specifically, for any j such that $\Theta^*_{j\cdot} \ne 0$, the first display in Section 2.1 yields $F^*_{j\cdot} = \Psi^*_{j\cdot} + B^{*T} A^*_{j\cdot}$. When $\Psi^*_{j\cdot}$ is sufficiently sparse, one could resort to the robust regression to estimate $\Psi^*_{j\cdot}$ (see details in [47]). Our procedure yields the index of non-zero rows of Ψ^* as well as the estimates of F^* and the row space of B^*. A full exploration of this approach is beyond the scope of this work and is left for future investigation.

In practice, we suggest to first check whether the conditions in Lemmas 3 and 4 are reasonable. If this is the case, our estimator $\widehat{\Theta}$ can be directly used for estimating Ψ^*. For example, the uniformity of U in Lemma 3 can be verified by comparing $\|\widehat{P}_{B^*} v\|_2^2$ with $\widehat{K}/m$ for some randomly generated unit vector $v \in \mathbb{R}^m$. Here $\widehat{P}_{B^*}$ is the estimate of P_{B^*} and $\widehat{K}$ is the estimated number of hidden variables. For conditions in Lemma 4, one could use Λ_K, the Kth eigenvalue of $B^{*T} \Sigma_W B^*$, as a surrogate of $\lambda_K(B^* B^{*T})$. The former can be estimated from the $\widehat{K}$th eigenvalue of the residual matrix $\widehat{\Sigma}_\varepsilon$, see Section 2.2.2 for details. Even if there is no prior information on the sparsity of rows of Ψ^*, Lemma 4 may still hold if $\widehat{K}$ is much smaller than m.

When conditions in Lemmas 3 and 4 seem questionable, we recommend to apply the procedure in (i) of **Case (2)** to check if $\widehat{\Theta}$ could be used to infer $\Psi^* C$ for some constraint matrix C with scientific interest. If this is not the case either, one may possibly apply the robust regression in (ii) of **Case (2)** to estimate certain rows of Ψ^*.

6. Simulation study.

In this section, we conduct simulations to verify our theoretical results. As mentioned in the Introduction, the SVA methods such as [37] require the condition $\Psi^* P_{B^*} \to 0$. To compare with [37] and other competing methods introduced below, we force $\Psi^* P_{B^*} = 0$ so that $\Theta^* = \Psi^*$ throughout this section.

Methods. We consider both HIVE and H-HIVE in Algorithms 1 and 3. All tuning parameters λ_1, λ_2 and λ_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 algorithm converges quickly in our simulation.

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

- Oracle: the estimator from (2.8) by using $P_B = B^T(B^TB)^{-1}B$ with the true B.
- Lasso: the group-lasso estimator from R-package glmnet.
- Ridge: the multivariate ridge estimator from R-package glmnet.
- HIVE-init: $\widehat{\Psi}$ obtained from solving (2.3) in step (1) of Algorithm 1.
- SVA: the surrogate variable analysis summarized in the following three steps: (i) compute $\widehat{\Theta}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$; (ii) obtain $\widehat{P}$ by the first K right singular vectors of $\boldsymbol{Y} - \boldsymbol{X}\widehat{\Theta}_{LS}$; (iii) estimate Θ^* by $\widehat{\Theta}_{LS}(\boldsymbol{I}_m - \widehat{P})$.[4]
- OLS: the ordinary least squares estimator $\widehat{\Theta}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

The Oracle estimator requires the knowledge of true B and is used as a benchmark to show the effect of estimating P_{B^*} on the estimation of Θ^* 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) (see more discussions in Remark 7).

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 [12] in Section 6.2.

Data generating mechanism. We set $K = s_* = 3$ throughout the simulation settings. The design matrix is sampled from $\boldsymbol{X}_{i\cdot} \sim N_p(0, \Sigma)$ for $1 \le i \le n$ where $\Sigma_{j\ell} = (-1)^{j+\ell}\rho^{|j-\ell|}$ for all $1 \le j, \ell \le p$. Under $Z = A^TX + W$, 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 $1 \le j \le p$, $1 \le k \le K$ and $1 \le \ell \le m$. We use η to control the magnitude of A hence the dense matrix $L = AB$. We generate the first s_* rows of Θ_{raw} by sampling each entry independently from $N(\mu_\Theta, \sigma_\Theta^2)$ and set the rest rows to 0. The final Θ is chosen as $\Theta_{raw}(\boldsymbol{I}_m - B^T(B^TB)^{-1}B)$ which has the same row sparsity as Θ_{raw} and satisfies $\Theta P_B = 0$. For the

[4] This procedure is based on [37]. There are other variants of SVA in the literature, for instance, [47, 41]. Since they have similar performances in our setting, we only consider the aforementioned one. A detailed comparison with other SVA-related procedures is given in Appendix E.

error terms, we independently generate $\boldsymbol{W}_{ik} \sim N(0,1)$ for all $1 \le i \le n$ and $1 \le k \le K$. For homoscedastic case, $\boldsymbol{E}_{ij}$ for $1 \le i \le n$ and $1 \le j \le m$ are i.i.d. realizations of $N(0,1)$. For heteroscedastic case, we independently generate $\boldsymbol{E}_{ij} \sim N(0, \tau_j^2)$, where, to vary the degree of heterogeneity, we follow the simulation setting in [50] and choose $\tau_j^2 = mv_j^\alpha / \sum_j v_j^\alpha$, where $v_1, \ldots, v_m$ are i.i.d. Unif[0, 1]. This choice of τ_j^2 guarantees $\sum_{j=1}^m \tau_j^2/m = 1$ and α controls the degree of heterogeneity: a larger α corresponds to more heterogeneity.

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 for the heteroscedastic case we vary $\alpha \in \{0, 3, 6, \ldots, 12, 15\}$ and fix $\eta = 0.5$. Within each combination of η and ρ (or α and ρ), we generate $\boldsymbol{X}$, A, B and Θ once and generate 100 replicates of the stochastic errors $\boldsymbol{W}$ and $\boldsymbol{E}$. For each method with their estimator $\widehat{\Theta}$ and the prediction $\boldsymbol{X}\widehat{F}$ (if available), we record the averaged Root Sum Squared Error (RSSE) $\|\widehat{\Theta} - \Theta\|_F$ and the averaged Prediction Mean Squared Error (PMSE) $\|\boldsymbol{X}\widehat{F} - \boldsymbol{X}F\|_F^2/(nm)$. We only report the results for $\rho = 0.5$ as the ones for $\rho = 0$ are similar.

6.1.1. *RSSE.*

The averaged RSSE 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 η increases such that the dense matrix L has larger magnitude. The reason is that the condition $Rem(P_{B^*}) \le c\sqrt{K}$ in Theorem 10 becomes restrictive for large p, small m and large η (say $\eta \ge 0.8$), 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 Θ^*. OLS is comparable to Ridge. Lasso has clear advantage over Ridge when the signal is sparse enough, that is, when η 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 η 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 7.

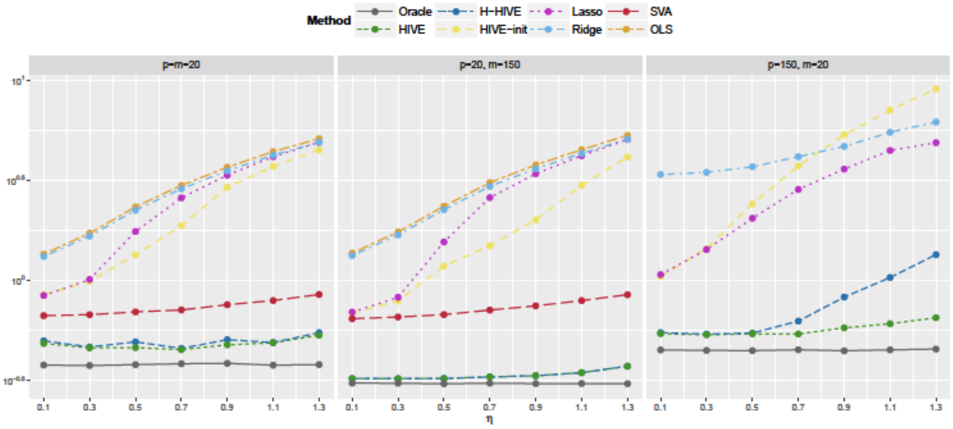

Fig 2: RSSE under the homoscedastic settings with $n = 100$.

Heteroscedastic cases: Figure 3 shows that H-HIVE, tailored for the heteroscedastic error, has the smallest RSSE 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 under heteroscedastic errors. 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 HeteroPCA. This is expected in light of Remark 8. Finally, Ridge, Lasso and HIVE-init are robust to the degree of heteroscedasticity in all cases, whereas SVA shows inflated RSSE as the degree of heteroscedasticity (α) increases when m is small (see the first panel).

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

Homoscedastic cases: As seen in the top row of Figure 4, when $n < p$ (the third panel), HIVE has much smaller PMSE than both Lasso and

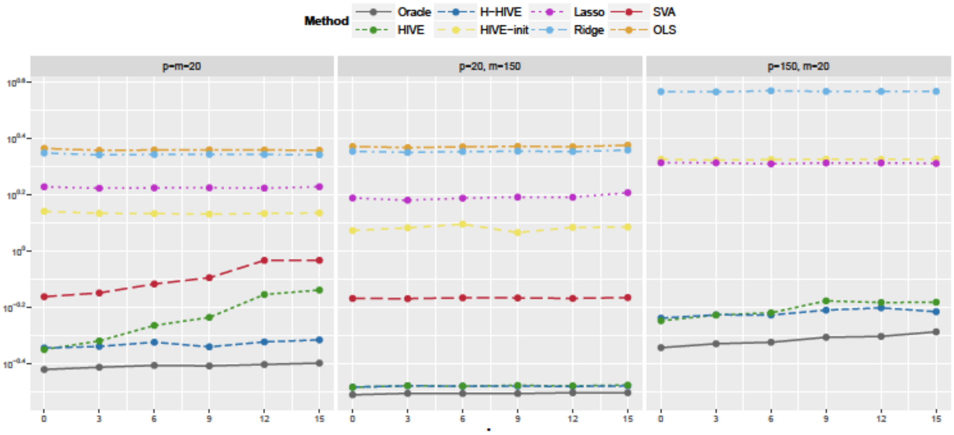

Fig 3: RSSE under the heteroscedastic settings with $n = 100$.

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 η (i.e. the signal $\Theta + L$ is approximately sparse). These findings are in line with Theorem 4 and its subsequent remarks.

Heteroscedastic cases: The bottom row of Figure 4 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 [12]. In the setting of $n = 100$, $m = 150$, $p = 20$, both methods select K consistently, which 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 $B_{kj} \sim N(0.1, 1)$. We choose $\Sigma_W = \sigma_W^2 \boldsymbol{I}_K$ 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 5 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

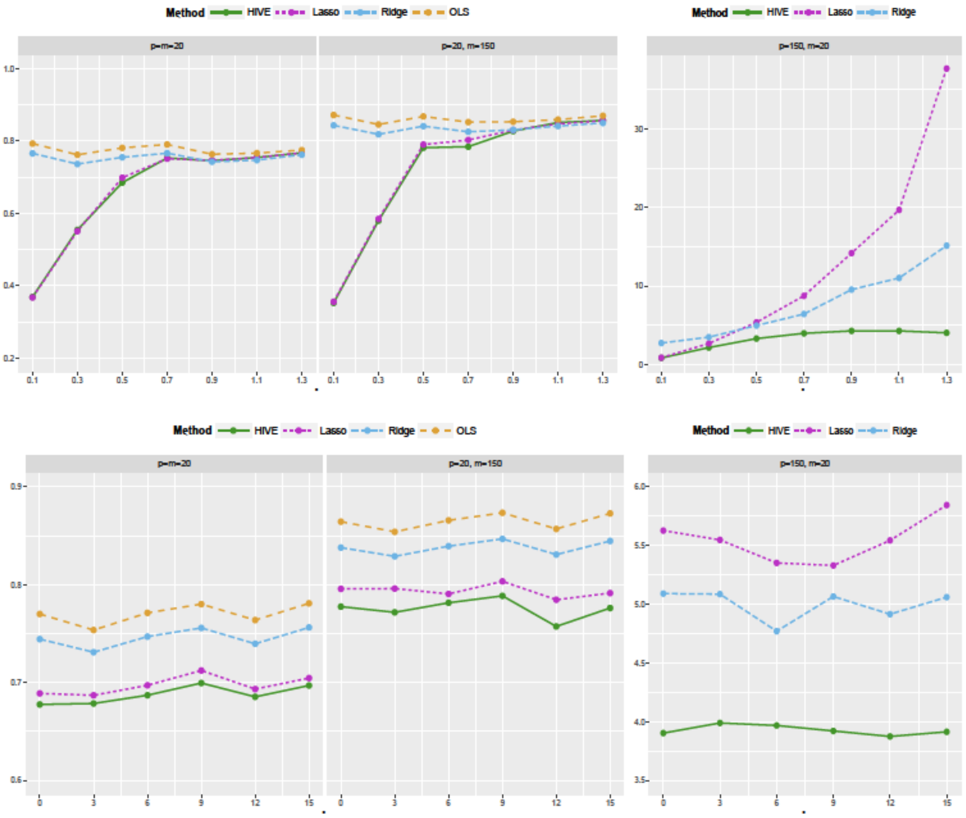

Fig 4: PMSE under the homoscedastic (the top row) and heteroscedastic (the bottom row) settings with $n = 100$.

to the implementation of SVD on the permuted data. For this reason, we recommend Ratio for moderate or large m.

7. Real data application.

We apply our procedures, Algorithms 1 and 3, to two real world 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 [35] 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 lo-

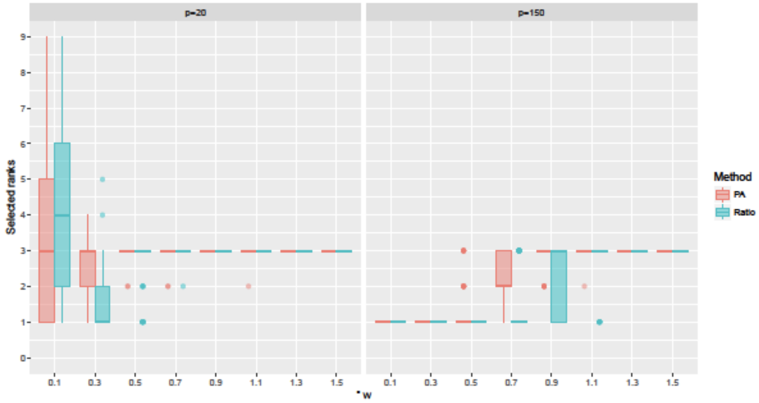

Fig 5: Boxplots of the selected K using PA and Ratio in two settings.

cation 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.

[13] showed that the data may exhibit a low-rank structure with estimated rank $K = 3$ via reduced-rank regression. This finding is consistent with [2] 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 [14] via the sparse reduced-rank regression (SSR) further reduces the LOOCV error to 304.5. Specifically, [14] estimated the coefficient matrix of the multivariate linear regression by

$$\widehat{F}_k = \min_{\text{rank}(F)\le k} \|\boldsymbol{Y} - \boldsymbol{X}F\|_F^2 + 2\lambda\|F\|_{\ell_1/\ell_2} \tag{7.1}$$

with $k = 3$ and $\lambda > 0$ selected from CV. The resulting estimator $\widehat{F}_k$ 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 $\widehat{K} = 3$. 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.

Table 1

LOOCV errors of HIVE, ridge, group-lasso (Lasso), SVA, reduced-rank regression (RRR) and sparse reduced-rank regression in (7.1) (SRR) on Norwegian dataset

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. [13] showed that the reduced-rank regression by using all p predictors can explain 86.9% of the total variation of $\boldsymbol{Y}$ quantified by $\operatorname{tr}(\boldsymbol{Y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y})$ (see [35] for the definition). Note that we can rewrite model (1.4) as a reduced-rank regression of $\boldsymbol{Y} - \boldsymbol{X}\Psi$ on $\boldsymbol{X}$. By replacing Ψ with our estimator $\widehat{\Psi}$, 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 ℓ_2 norm of rows of our estimator $\widehat{\Psi}+\widehat{L}$ corresponding to all the predictors. As a comparison, we also compute the reduced-rank estimator $\widetilde{L}$ by regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ directly. The results are also shown in Table 2. Similar to the results from the SRR, the estimator $\widetilde{L}$ corresponding to the three predictors X_1^2, X_1X_2 and X_2X_3 has small ℓ_2 norm. However, the association between the three predictors and responses is indeed strong as shown by our estimator $\widehat{\Psi}+\widehat{L}$ when the sparse signal $\widehat{\Psi}$ 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.

Table 2

ℓ_2 norms of rows of $\widetilde{L}$ and $\widehat{\Psi}+\widehat{L}$. The bold numbers correspond to the three excluded predictors, X_1^2, X_1X_2 and X_2X_3 in [14].

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

Yeast cross dataset. The yeast cross dataset[5] 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 con-

[5] The dataset is downloaded from http://genomics-pubs.princeton.edu/YeastCross_BYxRM/home.shtml.

sists 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 [10]. 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 [10]. 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 center and normalize the SNPs in the training set to zero mean and unit variance. Traits in the training set are also centered. The corresponding means and scales from the training set are used to standardize the test set. We then apply the HIVE Algorithm 1 together with group-lasso (Lasso) and Ridge to the training set and evaluate 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 refit the model to the whole dataset and apply 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 coefficient matrix Θ^*, which is adaptive to the unknown sparsity of Θ^*. The algorithm is further extended to settings 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, it is appealing 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. Second, it is of great interest to construct confidence intervals or hypothesis tests for the high-dimensional matrix Θ^* [30]. 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, our model (1.1) also covers a particular yet important model, the confounding model [18], which, in addition to (1.1), assumes $X = D'Z + W'$ for some random noise $W' \in \mathbb{R}^p$ independent of Z. As a result, the coefficient Ψ^* represents the causal effect of X on Y. Both our methods and analysis are directly applicable to this case. But it is also worth mentioning that assuming this extra structure brings the advantage of predicting Z by using X when p is large. The predicted Z in turn could potentially be used to improve the estimation of Ψ^*, especially when m is small. We do not pursue this direction in this paper and leave it to future research.

Acknowledgement. We thank the Associate Editor and two anonymous referees for their many insightful and helpful suggestions.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive Estimation of Multivariate Regression with Hidden Variables":
(doi: TBA). The supplementary document includes the proofs, the comparison with reduced-rank estimator and additional numerical results.

References.

[1] S. C. Ahn and A. R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3):1203–1227, 2013.

[2] M. Aldrin. Moderate projection pursuit regression for multivariate response data. *Computational Statistics & Data Analysis*, 21(5):501 – 531, 1996. ISSN 0167-9473. .

[3] T. W. Anderson. *An introduction to multivariate statistical analysis.* Wiley Series in Probability and Statistics. Wiley, 1984.

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

[5] J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

[6] Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.*, 21(3):1275–1294, 07 1993. .

[7] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788, 04 2014. .

[8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009. .

[9] X. Bing and M. H. Wegkamp. Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.*, 47(6):3157–3184, 12 2019. .

[10] J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237, 2013.

[11] P. Bühlmann and S. Van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[12] A. Buja and N. Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, 1992. . PMID: 26811132.

[13] F. Bunea, Y. She, and M. H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, 39(2):1282–1309, 04 2011. .

[14] F. Bunea, Y. She, and M. H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, 40(5):2359–2388, 10 2012. .

[15] F. Bunea, S. Strimas-Mackey, and M. Wegkamp. Interpolation under latent factor regression models, 2020.

[16] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.

[17] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, June 2011. ISSN 0004-5411. .

[18] D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral deconfounding via perturbed sparse linear models. *arXiv preprint arXiv:1811.05352*, 2018.

[19] S. Chakraborty, S. Datta, and S. Datta. Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics*, 28(6):799–806, 01 2012. ISSN 1367-4803. .

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

[21] V. Chernozhukov, C. Hansen, and Y. Liao. A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.*, 45(1):39–76, 02 2017. .

[22] E. Diaz. Causality and surrogate variable analysis. *arXiv:1704.00588*, 2017.

[23] L. H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 02 2016. .

[24] S. S. Dragomir. Reverses of schwarz inequality in inner product spaces with applications. *Mathematische Nachrichten*, 288(7):730–742, 2015. .

[25] J. Fan, Y. Liao, and M. Mincheva. High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, 39(6):3320–3356, 12 2011. .

[26] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680, 2013.

[27] J. Fan, L. Xue, and J. Yao. Sufficient forecasting using factor models. *Journal of Econometrics*, 201(2):292 – 306, 2017.

[28] J. A. Gagnon-Bartsch and T. P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 11 2012. .

[29] C. Giraud. Low rank multivariate regression. *Electron. J. Statist.*, 5:775–799, 2011. .

[30] Z. Guo, D. Ćevid, and P. Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding, 2020.

[31] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13(1):86, 2012. .

[32] J. J. Hox and T. M. Bechger. An introduction to structural equation modeling. 1998.

[33] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, Nov 2011. ISSN 1557-9654. .

[34] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, June 2014. ISSN 1615-3375. .

[35] A. Izenman. *Modern Multivariate. Statistical Techniques: Regression, Classification and Manifold Learning.* Springer, 2008.

[36] C. Lam and Q. Yao. Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.*, 40(2):694–726, 04 2012.

[37] S. Lee, W. Sun, F. A. Wright, and F. Zou. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika*, 104(2):303–316, 04 2017. ISSN 0006-3444. .

[38] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):1724–1735, 09 2007.

[39] J. T. Leek and J. D. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008. ISSN 0027-8424. .

[40] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 08 2011. .

[41] C. McKennan and D. Nicolae. Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika*, 106(4):823–840, 09 2019. ISSN 0006-3444. .

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

[43] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59(6):3434–3447, June 2013. ISSN 1557-9654. .

[44] Y. Sun, N. R. Zhang, and A. B. Owen. Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *Ann. Appl. Stat.*, 6(4):1664–1688, 12 2012. .

[45] A. E. Teschendorff, J. Zhuang, and M. Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 04 2011. ISSN 1367-4803. .

[46] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press, 2012. .

[47] J. Wang, Q. Zhao, T. Hastie, and A. B. Owen. Confounder adjustment in multiple hypothesis testing. *Ann. Statist.*, 45(5):1863–1894, 10 2017. .

[48] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 04 2014. ISSN 0006-3444. .

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

[50] A. Zhang, T. T. Cai, and Y. Wu. Heteroskedastic pca: Algorithm, optimality, and applications, 2018.

X. Bing, Y. Ning and Y. Xu
Department of Statistics and Data Science
Cornell University
Ithaca, New York 14853-3801
USA
E-mail: xb43@cornell.edu; yn265@cornell.edu; yx433@cornell.edu