Sparse CCA via Precision Adjusted Iterative Thresholding

Mengjie Chen

Computational Biology and Bioinformatics, Yale University, USA

Chao Gao

Department of Statistics, Yale University, USA

Zhao Ren

Department of Statistics, Yale University, USA

Harrison H. Zhou

Department of Statistics, Yale University, USA

Summary. Sparse Canonical Correlation Analysis (CCA) has received considerable attention in high-dimensional data analysis to study the relationship between two sets of random variables. However, there has been remarkably little theoretical statistical foundation on sparse CCA in high-dimensional settings despite active methodological and applied research activities. In this paper, we introduce an elementary sufficient and necessary characterization such that the solution of CCA is indeed sparse, propose a computationally efficient procedure, called CAPIT, to estimate the canonical directions, and show that the procedure is rate-optimal under various assumptions on nuisance parameters. The procedure is applied to a breast cancer dataset from The Cancer Genome Atlas project. We identify methylation probes that are associated with genes, which have been previously characterized as prognosis signatures of the metastasis of breast cancer.

Keywords: Canonical Correlation Analysis, Iterative Thresholding, Minimax Lower Bound, Optimal Convergence Rate, Single Canonical Pair Model, Sparsity

1. Introduction

Last decades witness the delivery of an incredible amount of information through the development of high-throughput technologies. Researchers now routinely collect a catalog of different measurements from the same group of samples. It is of great importance to elucidate the phenomenon in the complex system by inspecting the relationship between two or even more sets of measurements. Canonical correlation analysis is a popular tool to study

the relationship between two sets of variables. It has been successfully applied to a wide range of disciplines, including psychology and agriculture, and more recently, information retrieving (Hardoon et al., 2004), brain-computer interface (Bin et al., 2009), neuroimaging (Avants et al., 2010), genomics (Witten and Tibshirani, 2009) and organizational research (Bagozzi, 2011).

In this paper, we study canonical correlation analysis (CCA) in the high-dimensional setting. The CCA in the classical setting, a celebrated technique proposed by Hotelling (1936), is to find the linear combinations of two sets of random variables with maximal correlation. Given two centered random vectors $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$ with joint covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix}, \tag{1}$$

the population version of CCA solves

$$(\theta, \eta) = \arg\max_{(a,b)} \left\{ a^T \Sigma_{12} b : a^T \Sigma_1 a = 1, b^T \Sigma_2 b = 1 \right\}. \tag{2}$$

The optimization problem (2) can be solved by applying singular value decomposition (SVD) on the matrix $\Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}$. In practice, Hotelling (1936) proposed to replace $\Sigma_1^{-1/2} \Sigma_{12} \Sigma_2^{-1/2}$ by the sample version $\hat{\Sigma}_1^{-1/2} \hat{\Sigma}_{12} \hat{\Sigma}_2^{-1/2}$. This leads to consistent estimation of the canonical directions (θ, η) when the dimensions p_1 and p_2 are fixed and sample size n increases. However, in the high-dimensional setting, when the dimensions p_1 and p_2 are large compared with sample size n, this SVD approach may not work. In fact, when the dimensions exceed the sample size, SVD cannot be applied because the inverse of the sample covariance does not exist.

The difficulty motivates people to impose structural assumptions on the canonical directions in the CCA problem. For example, sparsity has been assumed on the canonical directions (Wiesel et al., 2008; Witten et al., 2009; Parkhomenko et al., 2009; Hardoon and Shawe-Taylor, 2011; Lê Cao et al., 2009; Waaijenborg and Zwinderman, 2009; Avants et al., 2010). The sparsity assumption implies that most of the correlation between two random vectors can be explained by only a small set of features or coordinates, which effectively reduces the dimensionality and at the same time improves the interpretability in many applications. However, to our best knowledge, there is no full characterization of the probabilistic CCA model that the canonical directions are indeed sparse. As a result, there has been remarkably little theoretical study on sparse CCA in high-dimensional settings despite recent active developments in methodology. This motivates us to find a sufficient and necessary condition on the covariance structure (1) such that the solution

of CCA is sparse. We show in Section 2 that (θ_1, η_1) is the solution of (2) if and only if (1) satisfies

$$\Sigma_{12} = \Sigma_1 \left(\sum_{i=1}^{r} \lambda_i \theta_i \eta_i^T \right) \Sigma_2, \tag{3}$$

where λ_i decreases $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_r > 0$, $r = \text{rank}(\Sigma_{12})$, and $\{(\theta_i, \eta_i)\}$ are orthonormal w.r.t. metric Σ_1 and Σ_2 respectively. i.e. $\theta_i^T \Sigma_1 \theta_j = \mathbb{I}\{i = j\}$ and $\eta_i^T \Sigma_2 \eta_j = \mathbb{I}\{i = j\}$. With this characterization, the canonical directions are sparse if and only if θ_1 and η_1 in (3) are sparse. Hence, sparsity assumption can be made explicit in this probabilistic model.

Motivated by the characterization (3), we propose a method called CAPIT, standing for Canonical correlation Analysis via Precision adjusted Iterative Thresholding, to estimate the sparse canonical directions. Our basic idea is simple. First, we obtain a good estimator of the precision matrices $(\Sigma_1^{-1}, \Sigma_2^{-1})$. Then, we transform the data by the estimated precision matrices to adjust the influence of the nuisance covariance (Σ_1, Σ_2). Finally, we apply iterative thresholding on the transformed data. The method is fast to implement in the sense that it achieves the optimal statistical accuracy in only finite steps of iterations.

Rates of convergence for the proposed estimating procedure are obtained under various sparsity assumptions on canonical directions and covariance assumptions on (Σ_1, Σ_2). In Section 4.2 we establish the minimax lower bound for the sparse CCA problem. The rates of convergence match the minimax lower bound as long as the estimation of nuisance parameters (Σ_1, Σ_2) is not dominating in estimation of the canonical directions. To the best of our knowledge, this is the first theoretically guaranteed method proposed in the sparse CCA literature.

We point out that the sparse CCA methods proposed in the literature may have both computational and statistical drawbacks. On the computational side, regularized versions of (2) such as Waaijenborg and Zwinderman (2009) and Wiesel et al. (2008) are proposed in the literature based on heuristics to avoid the non-convex nature of (2), but there is no theoretical guarantee whether these algorithms would lead to consistent estimators. On the statistical side, methods proposed in the literature do not explicitly take into account of the influence of the nuisance parameters. For example, Witten et al. (2009) and Parkhomenko et al. (2009) implicitly or explicitly use diagonal matrix or even identity matrix to approximate the unknown precision matrices $(\Sigma_1^{-1}, \Sigma_2^{-1})$. Such approximation could be valid when the covariance matrices (Σ_1, Σ_2) are nearly diagonal, otherwise there is no theoretical guarantee of consistency of the procedures. We illustrate this fact by a numerical example. We draw data from a multivariate Gaussian distribution and then
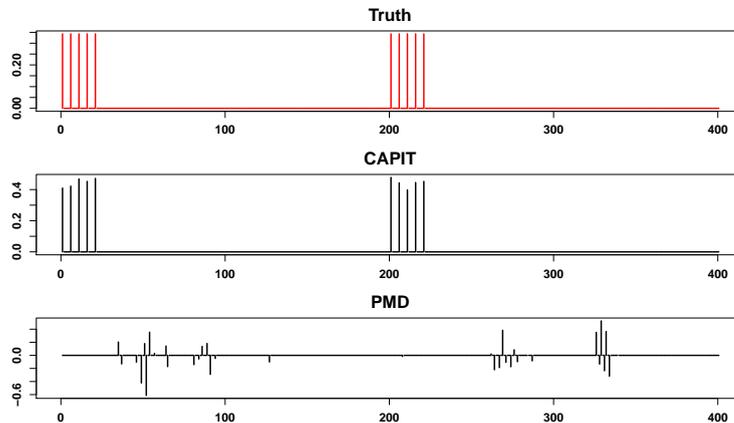
Fig. 1. Visualization of the simulation results of estimating (θ, η) for a replicate from Scenario II in Section 5.2 when $p = 200$ and $n = 500$. The $\{1, 2, ..., 200\}$-th coordinates represent θ and the $\{201, 202, ..., 400\}$-th coordinates represent η.

apply the proposed method and the Penalized Matrix Decomposition method by Witten et al. (2009). We show the results in Figure 1. By taking into account of the structure of the nuisance parameters, the CAPIT accurately recovers the sparse canonical directions, while the PMD is not consistent. In this simulation study, we consider sparse precision matrices and sparse canonical directions, where the sparse assumption of precision matrices has a sparse graphical interpretation of X and Y when the distribution is Gaussian. See Section 5.2 for more details.

A closely related problem is the principal component analysis (PCA) (Hotelling, 1933). In high-dimensional setting, sparse PCA is studied in Johnstone and Lu (2009), Ma (2013) and Cai et al. (2012). However, the PCA and CCA problems are fundamentally different. With the characterization of covariance structure in (3), such difference becomes clear. We illustrate the simplest rank-one case. Assuming the correlation rank r in (3) is one, the covariance structure is reduced to

$$\Sigma_{12} = \lambda \Sigma_1 \theta \eta^T \Sigma_2. \tag{4}$$

We refer to (4) as the Single Correlation Pair (SCP) model. In the PCA literature, the corresponding rank-one model is called the single-spike model. Its covariance structure can be written as

$$\Sigma = \lambda \theta \theta^T + I, \tag{5}$$

where θ is the principal direction of the random variable. A comparison of (4) and (5) reveals that estimation of the CCA is more involved than that of the PCA because of the

presence of the nuisance parameters (Σ_1, Σ_2), and the difficulty of estimating covariance matrices and precision matrices is known in high-dimensional statistics (Cai and Zhou, 2012; Ren et al., 2013). In the sparse PCA setting, the absence of nuisance parameter in (5) leads to algorithms directly applied on the sample covariance matrix $\hat{\Sigma}$, and the corresponding theoretical analysis is more tractable. In contrast, in the sparse CCA setting, not only do we need to adapt to the underlying sparsity of (θ, η), but we also need to adapt to the unknown covariance structure (Σ_1, Σ_2). We are going to show in Section 4 how various structures of (Σ_1, Σ_2) influence the convergence rate of the proposed method.

In addition, we demonstrate the CAPIT method by a real data example. We apply the proposed method to the data arising in the field of cancer genomics where methylation and gene expression are profiled for the same group of breast cancer patients. The method explicitly takes into account the sparse graphical model structure among genes. Interestingly, we identify methylation probes that are associated with genes that are previously characterized as prognosis signatures of the metastasis of breast cancer. This example suggests the proposed method provides a reasonable framework for exploratory and interpretive analysis of multiple datasets in high-dimensional settings.

The contributions in the paper are two-fold. First, we characterize the sparse CCA problem by proposing the probabilistic model and establish the minimax lower bound under certain sparsity class. Second, we propose the CAPIT method to adapt to both sparsity of the canonical direction and the nuisance structure. The CAPIT procedure is computationally efficient and attains optimal rate of convergence under various conditions. The paper is organized as follows. We first provide a full characterization of the sparse CCA model in Section 2. The CAPIT method and its associated algorithms are presented in Section 3. Section 4 is devoted to a theoretical analysis of our method. This section also presents the minimax lower bound. Section 5 and Section 6 investigate the numerical performance of our procedure by simulation studies and a real data example. The proof of the main theorem, Theorem 4.1, is gathered in Section 7. The proofs of all technical lemmas and Theorem 4.2 are gathered in Appendix.

1.1. Notations

For a matrix $A = (a_{ij})$, we use $||A||$ to denote its largest singular value and call it the spectral norm of A. The Frobenius norm is defined as $||A|| = \sqrt{\sum_{ij} a_{ij}^2}$. The matrix l_1 norm is defined as $||A||_{l_1} = \max_j \sum_i |a_{ij}|$. The norm $|| \cdot ||$, when applied to a vector, is understood as the usual Euclidean l_2 norm. For any two real numbers a and b, we use

notations $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Other notations will be introduced along with the text.

2. The Sparse CCA Model

Let $X \in \mathbb{R}^{p_1}$ and $Y \in \mathbb{R}^{p_2}$ be two centered multivariate random vectors with dimension p_1 and p_2 respectively. Write the covariance matrix of $(X^T, Y^T)^T$ as follows,

$$\mathrm{Cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix},$$

where Σ_1 is the covariance matrix of X with $\Sigma_1 = \mathbb{E}XX^T$, Σ_2 is the covariance matrix of Y with $\Sigma_2 = \mathbb{E}YY^T$, and $\Sigma_{21} = \Sigma_{12}^T$ the covariance structure between X and Y with $\Sigma_{12} = \Sigma_{21}^T = \mathbb{E}XY^T$. The canonical directions $\theta \in \mathbb{R}^{p_1}$ and $\eta \in \mathbb{R}^{p_2}$ are solutions of

$$\max_{a \in \mathbb{R}^{p_1}, b \in \mathbb{R}^{p_2}} \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_1 a}\sqrt{b^T \Sigma_2 b}}. \tag{6}$$

where we assume Σ_1 and Σ_2 are invertible and Σ_{12} is nonzero such that the maximization problem is not degenerate. Notice when (θ, η) is the solution of (6), $(\sigma_1 \theta, \sigma_2 \eta)$ is also the solution with arbitrary scalars (σ_1, σ_2) satisfying $\sigma_1 \sigma_2 > 0$. To achieve identifiability up to a sign, (6) can be reformulated into the following optimization problem.

$$\text{maximize } a^T \Sigma_{12} b, \quad \text{subject to } a^T \Sigma_1 a = 1 \quad \text{and} \quad b^T \Sigma_2 b = 1. \tag{7}$$

PROPOSITION 2.1. *When Σ_{12} is of rank 1, the solution (up to sign jointly) of Equation (7) is (θ, η) if and only if the covariance structure between X and Y can be written as*

$$\Sigma_{12} = \lambda \Sigma_1 \theta \eta^T \Sigma_2,$$

where $0 < \lambda \leq 1$, $\theta^T \Sigma_1 \theta = 1$ and $\eta^T \Sigma_2 \eta = 1$. In other words, the correlation between $a^T X$ and $b^T Y$ are maximized by $\mathrm{corr}(\theta^T X, \eta^T Y)$, and λ is the canonical correlation between X and Y.

The Proposition above is just an elementary consequence of SVD after transforming the parameters θ and η into $\Sigma_1^{1/2}\theta$ and $\Sigma_2^{1/2}\eta$ respectively. For the reasons of space, the proof is omitted. For general Σ_{12} with rank $r \geq 1$, it's a routine extension to see that the unique (up to sign jointly) solution of Equation (7) is (θ_1, η_1) if and only if the covariance structure between X and Y can be written as

$$\Sigma_{12} = \Sigma_1 \left(\sum_{i=1}^{r} \lambda_i \theta_i \eta_i^T \right) \Sigma_2,$$

where λ_i decreases $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_r > 0$, $\{\theta_i\}$ and $\{\eta_i\}$ are orthonormal w.r.t. metric Σ_1 and Σ_2 respectively. i.e. $\theta_i^T \Sigma_1 \theta_j = \mathbb{I}\{i = j\}$ and $\eta_i^T \Sigma_2 \eta_j = \mathbb{I}\{i = j\}$.

Inspired by (7), we propose a probabilistic model of (X, Y), so that the canonical directions (θ, η) are explicitly modeled in the joint distribution of (X, Y).

The Single Canonical Pair Model

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \lambda \Sigma_1 \theta \eta^T \Sigma_2 \\ \lambda \Sigma_2 \eta \theta^T \Sigma_1 & \Sigma_2 \end{pmatrix} \right), \tag{8}$$

with $\Sigma_1 \succ 0$, $\Sigma_2 \succ 0$, $\theta^T \Sigma_1 \theta = 1$, $\eta^T \Sigma_2 \eta = 1$ and $0 < \lambda \leq 1$.

Just as the single-spike model in PCA (Tipping and Bishop, 1999; Johnstone and Lu, 2009), the model (8) explicitly models (λ, θ, η) in the form of the joint distribution of (X, Y). Besides, it can be generalized to multiple canonical-pair structure as in the multi-spike model (Birnbaum et al., 2012). On the other hand, it is fundamentally different from the single-spike model, because (Σ_1, Σ_2) are typically unknown, so that estimating (θ, η) is much harder than estimating the spike in PCA. Even when both Σ_1 and Σ_2 are identity and $\Sigma_{12} = \lambda \theta \eta^T$, it cannot be reduced into the form of spike model. Bach and Jordan (2005) also proposed a statistical model for studying CCA in a probabilistic setting. Under their model, the data has a latent variable representation. It can be shown that the model we propose is equivalent to theirs in the sense that both can be written into the form of the other. The difference is that we explicitly model the canonical directions (θ, η) in the covariance structure for sparse CCA.

3. Methodology

In this section, we introduce the CAPIT algorithm to estimate the sparse canonical direction pair (θ, η) in the single canonical pair model in details. We start with the main part of the methodology in Section 3.1, an iterative thresholding algorithm, requiring an initializer and consistent estimators of precision matrices (nuisance parameters). Then in Section 3.2 we introduce a coordinate thresholding algorithm to provide a consistent initializer. Finally, in Section 3.3 rate-optimal estimators of precision matrices are reviewed over various settings.

The procedure is motivated by the power method, a standard technique to compute the leading eigenvector of a given symmetric matrix S (Golub and Van Loan, 1996). Let S be a $p \times p$ symmetric matrix. We compute its leading eigenvector. Starting with a vector

$v^{(0)}$ non-orthogonal to the leading eigenvector, the power method generates a sequence of vectors $v^{(i)}$, $i = 1, 2, \ldots$, by alternating the multiplication step $w^{(i)} = Sv^{(i-1)}$ and the normalization step $v^{(i)} = w^{(i)} / \|w^{(i)}\|$ until convergence. The limit of the sequence, denoted by $v^{(\infty)}$, is the leading eigenvector. The power method can be generalized to compute the leading singular vectors of any $p_1 \times p_2$ dimensional rectangular matrix M. Suppose the SVD of a rank d matrix M is $M = UDV^T$, where D is the d dimensional diagonal matrix with singular values on the diagonal. Suppose we are given an initial pair $(u^{(0)}, v^{(0)})$, non-orthogonal to the leading singular vectors. To compute the leading singular vectors, power method alternates the following steps until $(u^{(0)}, v^{(0)})$ converges to $(u^{(\infty)}, v^{(\infty)})$, which are the left and right leading singular vectors.

(a) Right Multiplication: $w_l^{(i)} = Mv^{(i-1)}$,
(b) Left Normalization: $u^{(i)} = w_l^{(i)} / \left\|w_l^{(i)}\right\|$,
(c) Left Multiplication: $w_r^{(i)} = u^{(i)}M$,
(d) Right Normalization: $v^{(i)} = w_r^{(i)} / \left\|w_r^{(i)}\right\|$.

Our goal is to estimate the canonical direction pair (θ, η). The power method above motivates us to find a matrix $\hat{A}$ close to $\lambda\theta\eta^T$ of which $(\theta/ \|\theta\| , \eta/ \|\eta\|)$ is the leading pair of singular vectors. Note that the covariance structure is $\Sigma_{12} = \lambda\Sigma_1\theta\eta^T\Sigma_2$. Suppose we know the marginal covariance structures of X and Y, i.e. $\Omega_1 = \Sigma_1^{-1}$ and $\Omega_2 = \Sigma_2^{-1}$ are given, it is very natural to consider $\Omega_1\hat{\Sigma}_{12}\Omega_2$ as the target matrix, where $\hat{\Sigma}_{12}$ is the sample cross-covariance between X and Y. Unfortunately, the covariance structures Ω_1 and Ω_2 are unknown as nuisance parameters, but a rate-optimal estimator of Ω_j $(j = 1, 2)$ usually can be obtained under various assumptions on the covariance or precision structures of X and Y in many high-dimensional settings. In literature, some commonly used structures are sparse precision matrix, sparse covariance matrix, bandable covariance matrix and Toeplitz covariance matrix structures. Later we will discuss the estimators of the precision matrices and their influences to the final estimation error of canonical direction pair (θ, η).

We consider the idea of data splitting. Suppose we have $2n$ i.i.d. copies $(X_i, Y_i)_{1 \leq i \leq 2n}$. We use the first half to compute the sample covariance $\hat{\Sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i^T$, and use the second half to estimate the precision matrices by $\hat{\Omega}_1$ and $\hat{\Omega}_2$. Hence the matrix $\hat{A} = \hat{\Omega}_1\hat{\Sigma}_{12}\hat{\Omega}_2$ is available to us. The reason for data splitting is that we can write the matrix $\hat{A}$ in an alternative form. That is,

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i^T,$$

where $\tilde{X}_i = \hat{\Omega}_1 X_i$ and $\tilde{Y}_i = \hat{\Omega}_2 Y_i$ for all $i = 1, ..., n$. Conditioning on $(X_{n+1}, Y_{n+1}), ..., (X_{2n}, Y_{2n})$, the transformed data $(\tilde{X}_i, \tilde{Y}_i)_{1 \leq i \leq n}$ are still independently identically distributed. This feature allows us to explore some useful concentration results in the matrix $\hat{A}$ to prove theoretical results. Conditioning on the second half of data, the expectation of $\hat{A}$ is $\lambda \alpha \beta^T$, where $\alpha = \hat{\Omega}_1 \Sigma_1 \theta$ and $\beta = \hat{\Omega}_2 \Sigma_2 \eta$. Therefore, the method we develop is targeted at (α, β) instead of (θ, η). However, as long as the estimators $(\hat{\Omega}_1, \hat{\Omega}_2)$ are accurate in the sense that

$$||\alpha - \theta|| \vee ||\beta - \eta|| = ||(\hat{\Omega}_1 \Sigma_1 - I)\theta|| \vee ||(\hat{\Omega}_2 \Sigma_2 - I)\eta||$$

is small, the final rate of convergence is also small.

If we naively apply the power method above to $\hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2$ in high-dimensional setting, the estimation variance accumulated across all p_1 and p_2 coordinates of left and right singular vectors goes very large and it is possible that we can never obtain a consistent estimator of the space spanned by the singular vectors. Johnstone and Lu (2009) proved that when $p/n \nrightarrow 0$, the leading eigenspace estimated directly from the sample covariance matrix can be nearly orthogonal to the truth under the PCA setting in which $\hat{A}$ is the sample covariance matrix with dimension $p_1 = p_2 = p$. Under the sparsity assumption of (θ, η), a natural way of reducing the estimation variance is to only estimate those coordinates with large values in θ and η respectively and simply estimate the rest coordinates by zero. Although bias is caused by this thresholding idea, in the end the variance reduction dominates the biased inflation and this trade-off minimizes the estimation error to the optimal rate. The idea of combining the power method and the iterative thresholding procedure leads to the algorithm in the next section which was also proposed by Yang et al. (2013) for a general data matrix $\hat{A}$ without a theoretical analysis.

3.1. Iterative Thresholding

We incorporate the thresholding idea into ordinary power method above for SVD by adding a thresholding step after each right and left multiplication steps before normalization. The thresholding step kills those coordinates with small magnitude to zero and keep or shrink the rest coordinates through a thresholding function $T(a, t)$ in which a is a vector and t is the thresholding level. In our theoretical analysis, we assume that $T(a, t) = \left(a_k \mathbb{I}\{|a_k| \geq t\} \right)$ is the hard-thresholding function, but any function serves the same purpose in theory as long as it satisfies (i) $|T(a, t)_k - a_k| \leq t$ and (ii) $T(a, t)_k = 0$ whenever $|a_k| < t$. Therefore the thresholding function can be hard-thresholding, soft-thresholding or SCAD

(Fan and Li, 2001). The algorithm is summarized below.

Algorithm 1: CAPIT: Iterative Thresholding

Input: Sample covariance matrices $\hat{\Sigma}_{12}$;

Estimators of precision matrix $\hat{\Omega}_1, \hat{\Omega}_2$;

Initialization pair $\alpha^{(0)}, \beta^{(0)}$;

Thresholding level γ_1, γ_2.

Output: Canonical direction estimator $\alpha^{(\infty)}, \beta^{(\infty)}$.

Set $\hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2$;

repeat

    Right Multiplication: $w^{l,(i)} = \hat{A}\beta^{(i-1)}$;

    Left Thresholding: $w^{l,(i)}_{th} = T\left(w^{l,(i)}, \gamma_1\right)$;

    Left Normalization: $\alpha^{(i)} = w^{l,(i)}_{th} / \left\|w^{l,(i)}_{th}\right\|$;

    Left Multiplication: $w^{r,(i)} = \alpha^{(i)}\hat{A}$;

    Right Thresholding: $w^{r,(i)}_{th} = T\left(w^{r,(i)}, \gamma_2\right)$;

    Right Normalization: $\beta^{(i)} = w^{r,(i)}_{th} / \left\|w^{r,(i)}_{th}\right\|$;

until *Convergence of* $\alpha^{(i)}$ *and* $\beta^{(i)}$;

REMARK 3.1. *In Algorithm 1, we don't provide specific stopping rule such as that the difference between successive iterations is small enough. For the single canonical pair model, we are able to show in Section 4 that the convergence is achieved in just one step. The intuition is simple: when A is of exact rank one, we can simply obtain the left singular vector via right multiplying A by any vector non-orthogonal to the right singular vector. Although in the current setting $\hat{A}$ is not a rank one matrix, the effect caused from the second singular value in nature does not change the statistical performance of our final estimator.*

REMARK 3.2. *The thresholding level (γ_1, γ_2) are user-specified. Theoretically, they should be set at the level $O\left(\sqrt{\frac{\log(p_1 \vee p_2)}{n}}\right)$. In Section 7.2, we present a fully data-driven (γ_1, γ_2) along with the proof.*

REMARK 3.3. *The estimator $(\alpha^{(\infty)}, \beta^{(\infty)})$ does not directly estimate (θ, η) because the former are unit vectors while the later satisfies $\theta^T \Sigma_1 \theta = \eta^T \Sigma_2 \eta = 1$. We are going to prove they are almost in the same direction by considering the loss function $|\sin \angle(a, b)|^2$ in Johnstone and Lu (2009). Details are presented in Section 4.*

REMARK 3.4. *The estimators of precision matrices $\hat{\Omega}_1$ and $\hat{\Omega}_2$ depend on the second half of the data and the estimator $\hat{\Sigma}_{12}$ depends on the first half of the data. In practice,*

after we apply Algorithm 1, we will swap the two parts of the data and use the first half to get $\hat{\Sigma}_{12}$ and the second half to obtain $\hat{\Omega}_1, \hat{\Omega}_2$. Then, Algorithm 1 is run again on the new estimators. The final estimator can be calculated through averaging the two. More generally, we can do sample splitting many times and take an average as Bagging, which is often used to improve the stability and accuracy of machine learning algorithms.

3.2. Initialization by Coordinate Thresholding

In Algorithm 1, we need to provide an initializer $\left(\alpha^{(0)}, \beta^{(0)}\right)$ as input. We generate a sensible initialization in this section which is similar to the "diagonal thresholding" sparse PCA method proposed by Johnstone and Lu (2009). Specifically, we apply a thresholding step to pick index sets B_1 and B_2 of the coordinates of θ and η respectively. Those index sets can be thought as strong signals. Then a standard SVD is applied on the submatrix of $\hat{A}$ with rows and columns indexed by B_1 and B_2. The dimension of this submatrix is relatively low such that the SVD on it is fairly accurate. The leading pair of singular vectors is of dimension $|B_1|$ and $|B_2|$, where $|\cdot|$ denotes the cardinality. In the end, we zero-pad the leading pair of singular vectors into dimension p_1 and p_2 respectively to provide our initializer $\left(\alpha^{(0)}, \beta^{(0)}\right)$. The algorithm is summarized in Algorithm 2.

Algorithm 2: CAPIT: Initialization by Coordinate Thresholding

Input: Sample covariance matrices $\hat{\Sigma}_{12}$;

Estimators of precision matrix $\hat{\Omega}_1, \hat{\Omega}_2$;

Thresholding level t_{ij}.

Output: Initializer $\alpha^{(0)}$ and $\beta^{(0)}$.

Set $\hat{A} = \hat{\Omega}_1 \hat{\Sigma}_{12} \hat{\Omega}_2$;

1 Coordinate selection: pick the index sets B_1 and B_2 of the coordinates of θ and η respectively as follows,
$$B_1 = \left\{i, \max_j |\hat{a}_{ij}| / t_{ij} \geq \sqrt{\tfrac{\log p_1}{n}}\right\}, B_2 = \left\{j, \max_i |\hat{a}_{ij}| / t_{ij} \geq \sqrt{\tfrac{\log p_2}{n}}\right\};$$

2 Reduced SVD: compute the leading pair of singular vectors $\left(\alpha^{(0),B_1}, \beta^{(0),B_2}\right)$ on the submatrix $\hat{A}_{B_1,B_2}$;

3 Zero-Padding procedure: construct the initializer $\left(\alpha^{(0)}, \beta^{(0)}\right)$ by zero-padding $\left(\alpha^{(0),B_1}, \beta^{(0),B_2}\right)$ on index sets B_1^c and B_2^c respectively,
$$\alpha_{B_1}^{(0)} = \alpha^{(0),B_1}, \alpha_{B_1^c}^{(0)} = 0, \beta_{B_2}^{(0)} = \beta^{(0),B_2}, \beta_{B_2^c}^{(0)} = 0.$$

The thresholding level t_{ij} in Algorithm 2 is a user specified constant and allowed to be adaptive to each location (i, j). The theoretical data-driven constant for each t_{ij} is provided in Section 7.2. It is clear the initializer is not unique since if $\left(\alpha^{(0)}, \beta^{(0)}\right)$ serves

as the output, $\left(-\alpha^{(0)}, -\beta^{(0)}\right)$ is also a solution of Algorithm 2. However either pair works as an initializer and provides the same result because in the end we estimate the space spanned by leading pair of singular vectors.

3.3. Precision Estimation

Algorithms 1 and 2 require precision estimators $\hat{\Omega}_1$ and $\hat{\Omega}_2$ to start with. As we mentioned, we apply the second half of the data to estimate the precision matrix $\hat{\Omega}_1$ and $\hat{\Omega}_2$. In this section, we discuss four commonly assumed covariance structures of X itself and provide corresponding estimators. We apply the same procedure to Y.

3.3.1. Sparse Precision Matrices

Precision matrix is closely connected to the undirected graphical model which is a powerful tool to model the relationships among a large number of random variables in a complex system. It is well known that recovering the structure of an undirected Gaussian graph is equivalent to recovering the support of the precision matrix. In this setting, it is natural to impose sparse graph structure among variables in X by assuming sparse precision matrices Ω_1. Many algorithms targeting on estimating sparse precision matrix were proposed in literature. See, e.g. Meinshausen and Bühlmann (2006), Friedman et al. (2008), Cai et al. (2011) and Ren et al. (2013). In the current paper, we apply the CLIME method to estimate Ω_1. For details of the algorithm, we refer to Cai et al. (2011).

3.3.2. Bandable Covariance Matrices

Motivated by applications in time series, where there is a natural "order" on the variables, the bandable class of covariance matrices was proposed by Bickel and Levina (2008a). In this setting, we assume that σ_{ij} decay to zero at certain rate as $|i - j|$ goes away from the diagonal. Usually regularizing the sample covariance matrix by banding or tapering procedures were applied in literature. We apply the tapering method proposed in Cai et al. (2010). Let $\omega = (\omega_m)_{0 \le m \le p-1}$ be a weight sequence with ω_m given by

$$\omega_m = \begin{cases} 1, & \text{when } m \le k/2 \\ 2 - \frac{2m}{k}, & \text{when } k/2 < m \le k \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

where k is the bandwidth. The tapering estimator $\hat{\Sigma}_1$ of the covariance matrix of X is given by $\hat{\Sigma}_1 = (\hat{\sigma}_{ij}^{sam} \omega_{|i-j|})$, where $\hat{\sigma}_{ij}^{sam}$ is the (i, j)-th entry of the sample covariance

matrix. The bandwidth k is chosen through cross-validation in practice. An alternative adaptive method was proposed by Cai and Yuan (2012). In the end, our estimator is $\hat{\Omega}_1 = \hat{\Sigma}_1^{-1}$.

3.3.3. *Toeplitz Covariance Matrices*

Toeplitz matrix is the symmetric matrix that the entries are constant along the off-diagonals which are parallel to the main diagonal. Class of Toeplitz covariance matrices arises naturally in the analysis of stationary stochastic processes. If X is a stationary process with autocovariance sequence $(\alpha_m) \equiv (\alpha_0, \alpha_1, \cdots, \alpha_{p-1}, \cdots)$, then the covariance matrix $\Sigma_1 = (\sigma_{ij})_{p_1 \times p_1}$ has a Toeplitz structure $\sigma_{ij} = \alpha_{|i-j|}$. In this setting, it is natural to assume certain rate of decay of the autocovariance sequence. We apply the following tapering method proposed in Cai et al. (2013). Define $\tilde{\sigma}_m = \frac{1}{p-m} \sum_{s-t=m} \hat{\sigma}_{st}^{sam}$, the average of sample covariance along each off-diagonal. Then the tapering estimator $\hat{\Sigma}_1 = (\hat{\sigma}_{st})$ with bandwidth k is defined as $\hat{\sigma}_{st} = \omega_{|s-t|}\tilde{\sigma}_{|s-t|}$, where $\omega = (\omega_m)_{0 \leq m \leq p-1}$ is defined in Equation (9). In practice, we pick bandwidth k using cross-validation. The final estimator of Ω_1 is then defined as $\hat{\Omega}_1 = \hat{\Sigma}_1^{-1}$.

3.3.4. *Sparse Covariance Matrices*

In many applications, there is no natural order on the variables like we assumed in bandable and Toeplitz covariance matrices. In this setting, permutation-invariant estimators are favored and general sparsity assumption is usually imposed on the whole covariance matrix, i.e. most of entries in each row/column of covariance matrix are zero or negligible. We apply a hard thresholding procedure proposed in Bickel and Levina (2008b) under this assumption. Again, let $\hat{\sigma}_{ij}^{sam}$ be the (i,j)-th entry of the sample covariance matrix of X. The thresholding estimator $\hat{\Sigma}_1 = (\hat{\sigma}_{st})$ is given by $\hat{\sigma}_{ij} = \hat{\sigma}_{ij}^{sam} I \left(|\hat{\sigma}_{ij}^{sam}| \geq \gamma \sqrt{\frac{\log p}{n}} \right)$ for some constant γ which is chosen through cross-validation. In the end, our estimator is $\hat{\Omega}_1 = \hat{\Sigma}_1^{-1}$.

4. Statistical Properties and Optimality

In this section, we present the statistical properties and optimality of our proposed estimator. We first present the convergence rates of our procedure, and then we provide a minimax lower bound for a wide range of parameter spaces. In the end, we can see when estimating the nuisance parameters is not harder than estimating canonical direction pair,

the rates of convergence match the minimax lower bounds. Hence we obtain the minimax rates of convergence for a range of sparse parameter spaces.

4.1. Convergence Rates

Notice that our model is fully determined by the parameter $(\Sigma_1, \Sigma_2, \lambda, \theta, \eta)$, among which we are interested in estimating (θ, η). To achieve statistical consistency, we need some assumptions on the interesting part (θ, η) and nuisance part $(\Sigma_1, \Sigma_2, \lambda)$.

Assumption A - Sparsity Condition on (θ, η):

We assume θ and η are in the weak l_q ball, with $0 \le q \le 2$. i.e.

$$\left|\theta_{(k)}\right|^q \le s_1 k^{-1}, \ \left|\eta_{(k)}\right|^q \le s_2 k^{-1},$$

where $\theta_{(k)}$ is the k-th largest coordinate by magnitude. Let $p = p_1 \vee p_2$ and $s = s_1 \vee s_2$. The sparsity levels s_1 and s_2 satisfy the following condition,

$$s = o\left(\left(\frac{n}{\log p}\right)^{\frac{1}{2} - \frac{q}{4}}\right). \tag{10}$$

REMARK 4.1. *In general, we can allow θ to be in the weak l_{q_1} ball and η to be in the weak l_{q_2} ball with $q_1 \ne q_2$. In that case, we require $s_i = o\left((n/\log p)^{\frac{1}{2} - \frac{q_i}{4}}\right)$ for $i = 1, 2$. There is no fundamental difference in the analysis and procedures. For simplicity, in the paper we only consider $q_1 = q_2$.*

Assumption B - General Conditions on $(\Sigma_1, \Sigma_2, \lambda)$:

(a) We assume there exist constants w and W, such that

$$0 < w \le \lambda_{\min}(\Sigma_i) \le \lambda_{\max}(\Sigma_i) \le W < \infty,$$

for $i = 1, 2$.

(b) In order that the signals do not vanish, we assume the canonical correlation is bounded below by a positive constant C_λ, i.e. $0 < C_\lambda \le \lambda \le 1$.

(c) Moreover, we require that estimators $(\hat{\Omega}_1, \hat{\Omega}_2)$ are consistent in the sense that

$$\xi_\Omega = ||\hat{\Omega}_1 \Sigma_1 - I|| \vee ||\hat{\Omega}_2 \Sigma_2 - I|| = o(1), \tag{11}$$

with probability at least $1 - O(p^{-2})$.

Loss Function

For two vectors a, b, a natural way to measure the discrepancy of their directions is the sin of the angle $|\sin \angle(a, b)|$, see Johnstone and Lu (2009). We consider the loss function

$L(a, b)^2 = 2|\sin \angle(a, b)|^2$. It is easy to calculate that

$$L(a, b) = \left\| \frac{aa^T}{||a||^2} - \frac{bb^T}{||b||^2} \right\|_F.$$

The convergence rate of the CAPIT procedure is presented in the following theorem.

THEOREM 4.1. *Assume the Assumptions A and B above hold. Let $(\alpha^{(k)}, \beta^{(k)})$ be the sequence from Algorithm 1, with the initializer $(\alpha^{(0)}, \beta^{(0)})$ calculated by Algorithm 2. The thresholding levels are*

$$t_{ij}, \quad \gamma_1 = c_1 \sqrt{\frac{\log p}{n}}, \quad \gamma_2 = c_2 \sqrt{\frac{\log p}{n}},$$

for sufficiently large constants (t_{ij}, c_1, c_2). Then with probability at least $1 - O(p^{-2})$, we have

$$L(\alpha^{(k)}, \theta)^2 \vee L(\beta^{(k)}, \eta)^2 \leq C \left(s \left(\frac{\log p}{n} \right)^{1-q/2} + ||(\hat{\Omega}_1 \Sigma_1 - I)\theta||^2 \vee ||(\hat{\Omega}_2 \Sigma_2 - I)\eta||^2 \right),$$

for all $k = 1, 2, ..., K$ with $K = O(1)$ and some constant $C > 0$.

REMARK 4.2. *Notice the thresholding levels depend on some unknown constants (t_{ij}, c_1, c_2). This is for the simplicity of presentation. A more involved fully data-driven choice of thresholding levels are presented in Section 7.2 along with the proof.*

The upper bound in Theorem 4.1 implies that the estimation of nuissance parameters $\hat{\Omega}_i$ affect the estimation canonical directions in terms of $||(\hat{\Omega}_1 \Sigma_1 - I)\theta||^2$ and $||(\hat{\Omega}_2 \Sigma_2 - I)\eta||^2$. In Section 3.3, we discussed four different settings in which certain structure assumptions are imposed on the nuance parameters Ω_1 and Ω_2. In the literature, optimal rates of convergence in estimating Ω_i under spectral norm have been established and can be applied here in each of the four settings, noting that $||(\hat{\Omega}_1 \Sigma_1 - I)\theta||^2 \leq ||(\hat{\Omega}_1 - \Omega_1)||^2 \, ||\Sigma_1 \theta||^2 \leq W ||(\hat{\Omega}_1 - \Omega_1)||^2$. Due to the limited space, we only discuss one setting in which we assume sparse precision matrix structure on Ω_i.

Besides the first general condition in Assumption B, we assume each row/column of Ω_i is in a weak l_{q_0} ball with $0 \leq q_0 \leq 1$. i.e. $\Omega_i \in \mathcal{G}_{q_0}(s_0, p_i)$ for $i = 1, 2$, where

$$\mathcal{G}_{q_0}(s_0, p) = \left\{ \Omega = (\omega_{ij})_{p \times p} : \max_j |\omega_{j(k)}|^{q_0} \leq s_0 k^{-1} \text{for all } k \right\},$$

and the matrix l_1 norm of Ω_i is bounded by some constant $||\Omega_i||_{l_1} \leq w^{-1}$. The notation $\omega_{j(k)}$ means the k-th largest coordinate of j-th row of Ω in magnitude. Recall that $p = p_1 \vee p_2$. Under the assumptions that $s_0^2 = O\left((n/\log p)^{1-q_0}\right)$, Theorem 2 in Cai et al.

(2011) implies that CLIME estimator with an appropriate tuning parameter attaining the following rate of convergence $||(\hat{\Omega}_1\Sigma_1 - I)\theta||^2$ with probability at least $1 - O\left(p^{-2}\right)$,

$$||(\hat{\Omega}_1\Sigma_1 - I)\theta||^2 \leq W||(\hat{\Omega}_1 - \Omega_1)||^2 \leq Cs_0^2 \left(\frac{\log p}{n}\right)^{1-q_0}.$$

Therefore we obtain the following corollary.

COROLLARY 4.1. *Assume the Assumptions A and B holds, $\Omega_i \in \mathcal{G}_{q_0}(s_0, p_i)$ $i = 1, 2$, $\|\Omega_i\|_{l_1} \leq w^{-1}$ and $s_0^2 = O\left((n/\log p)^{1-q_0}\right)$. Let $(\alpha^{(k)}, \beta^{(k)})$ be the sequence from Algorithm 1, with the initializer $(\alpha^{(0)}, \beta^{(0)})$ calculated by Algorithm 2 and $\hat{\Omega}_i$ obtained by applying CLIME procedure in Cai et al. (2011). The thresholding levels are the same as those in Theorem 4.1. Then with probability at least $1 - O(p^{-2})$, we have*

$$L(\alpha^{(k)}, \theta)^2 \vee L(\beta^{(k)}, \eta)^2 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + s_0^2\left(\frac{\log p}{n}\right)^{1-q_0}\right),$$

for all $k = 1, 2, ..., K$ with $K = O(1)$ and some constant $C > 0$.

REMARK 4.3. *It can be seen from the analysis that similar upper bounds hold in Corollary 4.1 with probability $1 - O(p^{-h})$ by picking different thresholding constants in Algorithms 1, 2 and CLIME procedure for any $h > 0$. Assuming that $n = o(p^h)$, the boundedness of loss function implies that Corollary 4.1 is valid in the risk sense.*

4.2. Minimax Lower Bound

In this section, we establish a minimax lower bound in a simpler setting in which we know the covariance matrices Σ_1 and Σ_2. We assume $\Sigma_i = I_{p_i \times p_i}$ for $i = 1, 2$ for simplicity. Otherwise, we can transfer the data accordingly and make $\Sigma_i = I_{p_i \times p_i}$. The purpose of establishing this minimax lower bound is to measure the difficulty of estimation problems in sparse CCA model. In view of the upper bound given in Theorem 4.1 by the iterative thresholding procedure Algorithms 1 and 2, this lower bound is minimax rate optimal under conditions that estimating nuisance precision matrices is not harder than estimating the canonical direction pair. Consequently, assuming some general structures on the nuance parameters Σ_1 and Σ_2, we establish the minimax rates of convergence for estimating the canonical directions.

Before proceeding to the precise statements, we introduce the parameter space of (θ, η, λ) in this simpler setting. Define

$$\mathcal{F}_q^{p_1, p_2}(s_1, s_2, C_\lambda) = \left\{ \begin{array}{c} N(0, \Sigma) : \Sigma \text{ is specified in } (8), \lambda \in (C_\lambda, 1) \\ \Sigma_i = I_{p_i \times p_i}, i = 1, 2, \\ |\theta|_{(k)}^q \leq s_1 k^{-1}, \ |\eta|_{(k)}^q \leq s_2 k^{-1}, \text{for all } k. \end{array} \right\}. \tag{12}$$

In the sparsity class (12), the covariance matrices $\Sigma_i = I_{p_i \times p_i}$ for $i = 1, 2$ are known and unit vectors θ, η are in the weak l_q ball, with $0 \le q \le 2$. We allow the dimensions of two random vectors p_1 and p_2 to be very different and only require that $\log p_1$ and $\log p_2$ are comparable with each other,

$$\log p_1 \asymp \log p_2. \tag{13}$$

Remember $s = s_1 \vee s_2$ and $p = p_1 \vee p_2$.

THEOREM 4.2. *For any $q \in [0, 2]$, we assume that $s_i \left(\frac{n}{\log p_i} \right)^{q/2} = o(p_i)$ for $i = 1, 2$ and (13) holds. Moreover, we also assume $s \left(\frac{\log p}{n} \right)^{1 - \frac{q}{2}} \le c_0$, for some constant $c_0 > 0$. Then we have*

$$\inf_{(\hat{\theta}, \hat{\eta})} \sup_{P \in \mathcal{F}} \mathbb{E}_P \left(L^2(\hat{\theta}, \theta) \vee L^2(\hat{\eta}, \eta) \right) \ge C s \left(\frac{\log p}{n} \right)^{1 - q/2},$$

where $\mathcal{F} = \mathcal{F}_q^{p_1, p_2} (s_1, s_2, C_\lambda)$ and C is a constant only depending on q and C_λ.

Theorem 4.2 implies the minimaxity for the sparse CCA problem when the covariance matrices Σ_1 and Σ_2 are unknown. The lower bound directly follows from Theorem 4.2 and the upper bound follows from Theorem 4.1. Define the parameter space

$$\mathcal{F}_{q,q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W) = \left\{ \begin{array}{c} N(0, \Sigma) : \Sigma \text{ is specified in } (8), \lambda \in (C_\lambda, 1), \\ \Sigma_i^{-1} \in \mathcal{G}_{q_0} (s_0, p_i), W^{-1} \le \lambda_{\min}(\Sigma_i^{-1}), \left\| \Sigma_i^{-1} \right\|_{l_1} \le w^{-1}, \\ |\theta|_{(k)} \le s_1 k^{-1}, \; |\eta|_{(k)} \le s_2 k^{-1} \text{for all } k. \end{array} \right\}.$$

Since $\mathcal{F}_q^{p_1, p_2} (s_1, s_2, C_\lambda) \subset \mathcal{F}_{q,q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W)$, the lower bound for the smaller space holds for the larger one. Combining the Corollary 4.1 and the minimax lower bound in Theorem 4.2, we obtain that the minimax rate of convergence of estimating canonical directions over parameter spaces $\mathcal{F}_{q,q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W)$.

COROLLARY 4.2. *Under the assumptions in Corollary 4.1 and Theorem 4.2 and assume $n = o(p^h)$ for some $h > 0$, we have*

$$\inf_{(\hat{\theta}, \hat{\eta})} \sup_{P \in \mathcal{F}} \mathbb{E}_P \left(L(\hat{\theta}, \theta)^2 \vee L(\hat{\eta}, \eta)^2 \right) \asymp s \left(\frac{\log p}{n} \right)^{1 - q/2},$$

for $\mathcal{F} = \mathcal{F}_{q,q_0}^{p_1, p_2} (s_0, s_1, s_2, C_\lambda, w, W)$, provided that $s_0^2 \left(\frac{\log p}{n} \right)^{1 - q_0} \le C s \left(\frac{\log p}{n} \right)^{1 - q/2}$ for some constant $C > 0$.

5. Simulation Studies

We present simulation results of our proposed method in this section. In the first scenario, we assume the covariance structure is sparse, and in the second scenario, we assume the

precision structure is sparse. Comments on both scenarios are addressed at the end of the section.

5.1. Scenario I: Sparse Covariance Matrix

In the first scenario, we consider covariance matrices Σ_1 and Σ_2 are sparse. More specifically, the covariance matrix $\Sigma_1 = \Sigma_2 = (\sigma_{ij})_{1 \leq i,j \leq p}$ takes the form

$$\sigma_{ij} = \rho^{|i-j|} \quad \text{with} \quad \rho = 0.3.$$

The canonical pair (θ, η) is generated by normalizing a vector taking the same value at the coordinates $(1, 6, 11, 16, 21)$ and zero elsewhere such that $\theta^T \Sigma_1 \theta = 1$ and $\eta^T \Sigma_2 \eta = 1$. The canonical correlation λ is taken as 0.9. We generate the $2n \times p$ data matrices X and Y jointly from (8). As described in the methodology section, we split the data into two halves. In the first step, we estimate the precision matrices Ω_1 and Ω_2 using the first half of the data. Note that this covariance matrix has a Toeplitz structure. We estimate the covariance matrix under three different assumptions: 1) we assume that the Toeplitz structure is known and estimate $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ by the method proposed in Cai et al. (2013) (denoted as CAPIT+Toep); 2) we assume that it is known that covariance σ_{ij} decay as they move away from the diagonal and estimate $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ by the tapering procedure proposed in Cai et al. (2010) (denoted as CAPIT+Tap); 3) we assume only the sparse structure is known and estimate $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ by hard thresholding (Bickel and Levina, 2008b) (denoted as CAPIT+Thresh). In the end the estimators $\hat{\Omega}_i$ is given by $\hat{\Omega}_i = \hat{\Sigma}_i^{-1}$ for $i = 1, 2$.

To select the tuning parameters for different procedures, we further split the first part of the data into a $2 : 1$ training set and tuning set. We select the tuning parameters by minimizing the distance of estimated covariance from the training set and sample covariance matrix of the tuning set in term of the Frobenius norm. More specifically, the tuning parameters k_1 in the Toeplitz method and k_2 in the Tapering method are selected through a screening on numbers in the interval of $(1, p)$. The tuning parameter λ in the Thresholding method is selected through a screening on 50 numbers in the interval of $[0.01, 0.5]$.

After obtaining estimator $\hat{\Omega}_1$ and $\hat{\Omega}_2$, we perform Algorithms 1 and 2 by using $\hat{\Sigma}_{12}$ estimated from the second half of the data. The thresholding parameters γ_1 and γ_2 are set to be $2.5\sqrt{\frac{\log p}{n}}$ for the Tapering and Thresholding methods, while the thresholding parameter t_{ij} is set to be 2.5 for all (i, j). For the Toeplitz method, the thresholding parameters $\gamma_1 = \gamma_2 = 2\sqrt{\frac{\log p}{n}}$ while parameter $t_{ij} = 2$ for all (i, j). The resulted estimator

Table 1. Scenario I: Sparse covariance matrix. Estimation errors for (θ, η) as measured by $L(\hat{\theta}, \theta) \vee L(\hat{\eta}, \eta)$ based on the median of 100 replications. Numbers in parentheses are the simulation median absolute deviations.

$p_1 = p_2$	n	CAPIT+Toep	CAPIT+Tap	CAPIT+Thresh	PMD	SVD
200	750	0.11(0.03)	0.12(0.06)	0.11(0.03)	0.16(0.03)	0.32(0.01)
300	750	0.11(0.03)	0.13(0.07)	0.11(0.03)	0.36(0.02)	0.44(0.01)
200	1000	0.1(0.02)	0.1(0.05)	0.09(0.03)	0.14(0.02)	0.27(0.01)
500	1000	0.09(0.03)	0.09(0.04)	0.1(0.02)	0.11(0.03)	0.53(0.02)

is denoted as $(\hat{\theta}_{[1]}, \hat{\eta}_{[1]})$.

Then we swap the data, repeat the above procedures and obtain $(\hat{\theta}_{[2]}, \hat{\eta}_{[2]})$. The final estimator $(\hat{\theta}, \hat{\eta})$ is the average of $(\hat{\theta}_{[1]}, \hat{\eta}_{[1]})$ and $(\hat{\theta}_{[2]}, \hat{\eta}_{[2]})$.

We compare our method with penalized matrix decomposition proposed by Witten et al. (2009) (denoted as PMD) and the vanilla singular vector decomposition method for CCA (denoted as SVD). For PMD, we use the R function implemented by the authors (Witten et al., 2013), which performs sparse CCA by l_1-penalized matrix decomposition and selects the tuning parameters using a permutation scheme.

We evaluate the performance of different methods by the loss function $L(\hat{\theta}, \theta) \vee L(\hat{\eta}, \eta)$. The results from 100 independent replicates are summarized in Table 1.

5.2. *Scenario II: Sparse Precision Matrix*

In the second scenario, we consider that the precision matrices Ω_1 and Ω_2 are sparse. In particular, $\Omega_1 = \Omega_2 = (\omega_{ij})_{1 \le i,j \le p}$ take the form:

$$\omega_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0.5 & \text{if } |i - j| = 1 \\ 0.4 & \text{if } |i - j| = 2 \\ 0 & \text{otherwise.} \end{cases}$$

The canonical pair (θ, η) is the same as described in Scenario I. We generate the $2n \times p$ data matrices X and Y jointly from (8).

As described in the methodology section, we split the data into two halves. In the first step, we estimate the precision matrices by the CLIME proposed in Cai et al. (2011) (denoted as CAPIT+CLIME). The tuning parameter λ is selected by maximizing the log-likelihood function. In the second step, we perform Algorithms 1 and 2 with $\hat{\Sigma}_{12}$ estimated

Table 2. Scenario II: Sparse precision matrix. Estimation errors for (η, θ) as measured by $L(\hat{\theta}, \theta) \vee L(\hat{\eta}, \eta)$ based on the median of 100 replications. Numbers in parentheses are the simulation median absolute deviations.

$p_1 = p_2$	n	CAPIT+CLIME	PMD	SVD
200	500	0.41(0.35)	1.41(0)	0.52(0.03)
200	750	0.2(0.05)	1.19(0.33)	0.39(0.02)
500	750	0.21(0.12)	1.41(0)	0.84(0.03)

from the second half. The thresholding parameter γ_1 and γ_2 are set to be $1.5\sqrt{\frac{\log p}{n}}$ and t_{ij} is set to be 1.5. The resulted estimator is denoted as $(\hat{\theta}_{[1]}, \hat{\eta}_{[1]})$. Then we swap the data, repeat the above procedures and obtain $(\hat{\theta}_{[2]}, \hat{\eta}_{[2]})$. The final estimator $(\hat{\theta}, \hat{\eta})$ is the average of $(\hat{\theta}_{[1]}, \hat{\eta}_{[1]})$ and $(\hat{\theta}_{[2]}, \hat{\eta}_{[2]})$.

For comparison, we also apply PMD and SVD in this case. The results from 100 independent replicates are summarized in Table 2. A visualization of the estimation from a replicate in from the case $n = 500, p = 200$ under Scenario II is shown in Figure 1.

5.3. *Discussion on the Simulation Results*

The above results (Table 1 and Table 2) show that our method outperforms the PMD method proposed by Witten et al. (2009) and the vanilla SVD method (Hotelling, 1936). It is not surprising that the SVD method does not perform better than our method because of the sparse assumption in the signals. We focus our discussion on the comparison of our method and the PMD method.

The PMD method is defined by the solution of the following optimization problem

$$(\hat{\theta}_{PMD}, \hat{\eta}_{PMD}) = \arg\max_{(u,v)} \left\{ u^T \hat{\Sigma}_{12} v : ||u|| \leq 1, ||v|| \leq 1, ||u||_1 \leq c_1, ||v||_1 \leq c_2 \right\}.$$

As noted by Witten et al. (2009), the PMD method approximates the covariance Σ_1 and Σ_2 by the identity matrices $I_{p_1 \times p_1}$ and $I_{p_2 \times p_2}$. If we ignore the l_1 regularization, the population version of PMD is to maximize $u^T \Sigma_{12} v$ subject to $||u|| \vee ||v|| \leq 1$, which gives the maximizer in the direction of $(\Sigma_1 \theta, \Sigma_2 \eta)$ instead of (θ, η). When the covariance matrices Σ_1 and Σ_2 are sufficiently sparse, $(\Sigma_1 \theta, \Sigma_2 \eta)$ and (θ, η) are close. This explains that in Scenario I, the PMD method performs well. However, in Scenario II, we assume the precision matrices Ω_1 and Ω_2 are sparse. In this case, the corresponding Σ_1 and Σ_2 are not necessarily sparse, implying that $(\Sigma_1 \theta, \Sigma_2 \eta)$ could be far away from (θ, η). The PMD method is not consistent in this case, as is illustrated in Figure 1. In contrast, our

method takes advantage of the sparsity of Ω_1 and Ω_2, and accurately recovers the canonical directions.

6.   Real Data Analysis

DNA methylation plays an essential role in the transcriptional regulation (VanderKraats et al., 2013). In tumor, DNA methylation patterns are frequently altered. However, how these alterations contribute to the tumorigenesis and how they affect gene expression and patient survival remain poorly characterized. Thus it is of great interest to investigate the relationship between methylation and gene expression and their interplay with survival status of cancer patients. We applied the proposed method to a breast cancer dataset from The Cancer Genome Atlas project (TCGA, 2012). This dataset consists both DNA methylation and gene expression data for 193 breast cancer patients. The DNA methylation was measured from Illumina Human methylation 450 BeadChip, which contains 482,431 CpG sites that cover 96% of the genome-wide CpG islands. Since no batch effect has either been reported from previous studies or been observed from our analysis, we do not further process the data. For methylation data, there are two popular metrics used to measure methylation levels, β-value and M-value statistics. β-value is defined as the proportion of methylated probes at a CpG site. M-value is defined as the $\log 2$ ratio of the intensities of methylated probe versus un-methylated probe, which is reported as approximately homoscedastic in a previous study Du et al. (2010). We choose to use M-value for methylation data in our analysis.

To investigate the relationship of methylation and gene expression and their interplay with clinical outcomes, we follow the supervised sparse CCA procedure suggested in Witten and Tibshirani (2009). More specifically, we first select methylation probes and genes that are marginally associated with the disease free status by performing a screening on methylation and gene expression data, respectively. There are 135 genes and 4907 methylation probes marginally associated with disease free status with a P-value less than 0.01. We further reduce the number of methylation probes to 3206 by selecting the ones with sample variance greater than 0.5. Compared to the sample size, the number of methylation probes is still too large. To control the dimension of input data, we apply our methods to 135 genes with the methylation probes on each chromosome separately. Since it is widely believed that genes operate in biological pathways, the graph for gene expression data is expected to be sparse. We apply the proposed procedure under the sparse precision matrix setting (Section 3.3.1). As we have discussed in the simulation studies, the canonical
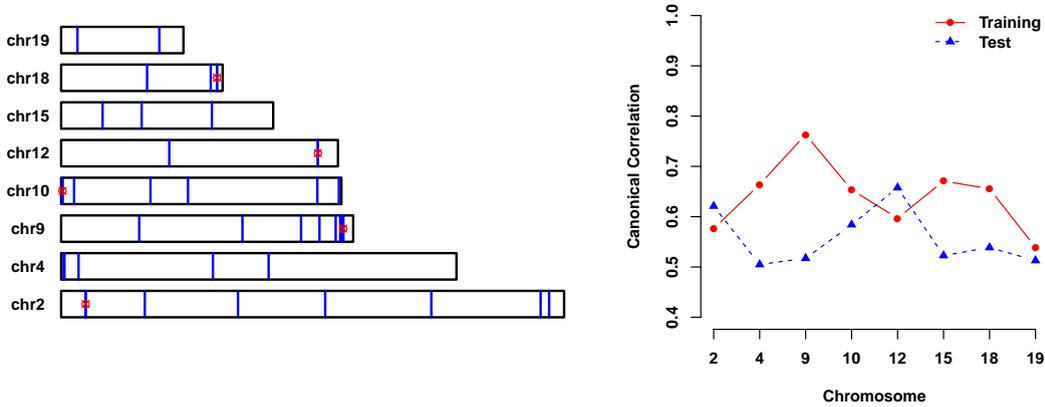
Fig. 2. Left: Visualization of the genomic coordinates of detected methylation probes. Genes that represented by more than one probes are highlighted by square symbols. Right: Canonical correlations of disease associated genes and methylation probes on eight chromosomes in the training set and the test set.

correlation structure under the sparse precision matrix setting cannot be estimated by the current methods in the literature, such as PMD.

For the purpose of interpretation, the tuning parameters are selected such that a sparse representation of $(\hat{\theta}, \hat{\eta})$ is obtained while the canonical correlation is high. More specifically, we require the number of non-zero genes or probes is less than 10 for each chromosome. We split the data into two halves as a test set and a training set. Then we applied the proposed procedure on the training set. To remove false discoveries, we required the canonical correlation on the test set is greater than 0.5. In total, there are eight chromosomes that have methylation probes satisfying the above criteria (shown in Figure 2). In Table 3, we list genes and methylation probes on each chromosome that form the support of detected canonical directions. A further examination of the genomic coordinates of detected methylation probes reveal the physical closeness of some probes. Some detected probes correspond to the same gene. LPIN1 on chromosome 2, RXRA on chromosome 9, DIP2C on chromosome 10, AACS on chromosome 12, and NFATC1 on chromosome 18 are represented by more than one methylation probes (shown in Figure 2). Moreover, 16 of the 25 genes listed in Table 3 are detected more than once as candidate genes associated with methylation probes. ORC6L, RRM2, RAB6B are independently detected from four

Table 3. Sparse CCA results for methylation sites and gene expression that are associated with disease free status for TCGA breast cancer data. In the analysis, methylation and gene expression data are assumed to have sparse precision matrix structure. Sparse CCA were performed on the same set of genes with methylation probes on different chromosomes. Chromosomes with canonical correlation greater than 0.5 on both the training and test set are listed with the number of probes associated with disease free status and the probes form the support of canonical directions.

Chromosome	Number of probes	Genes
		Methylation probes
2	269	RRM2, ILF2, ORC6L, SUSD3, SHCBP1
		cg04799980, cg08022717, cg10142874, cg13052887, cg16297938, cg24011073, cg26364080, cg27060355
4	143	SLC26A9, C15orf52, NPM2, DNAH11, RAB6B, LIN28, STC2
		cg04812351, cg14505741, cg15566751, cg15763121, cg17232991
9	89	RGS6, ORC6L, PTPRH, GPX2, QSOX2, NPM2, SCG3, RAB6B, L1CAM, STC2, REG1A
		cg01729066, cg02127980, cg03693099, cg13413384, cg13486627, cg13847987, cg14004457, cg14443041, cg21123355
10	116	RRM2, SLC26A9, ORC6L, PTPRH, DNAH11, SCG3, LIN28, UMODL1, C11orf9
		cg00827318, cg01162610, cg01520297, cg03182620, cg14522790, cg14999931, cg19302462
12	175	QSOX2, SCG3
		cg00417147, cg13074795, cg21881338
15	92	C15orf52, NPM2, DNAH11, SELE, RAB6B
		cg11465404, cg18581777, cg21735516
18	37	ILF2, SPRR2D, ADCY4, RAB6B, C11orf9, REG1A, SHCBP1
		cg07740306, cg15531009, cg18935516, cg19363889
19	162	ORC6L, NPM2, GPR56
		cg06392698, cg06555246

Table 4. Detected methylation probes and their corresponding genes on chromosome 9.

Probe	Gene	Function	η_i
cg01729066	MIR600	microRNA regulating estrogen factors	0.282
cg14004457	MIR455	microRNA regulating estrogen factors	0.347
cg02127980, cg13413384	RXRA	retinoic X receptors	0.269, 0.242
cg03693099	CEL	fat catalyzation and vitamin absorption	0.286
cg13486627	RG9MTD3	RNA (guanine-9-) methyltransferase	-0.479
cg13847987	ABL1	a protein tyrosine kinase functioned in cell differentiation and stress response	0.334
cg21123355	VAV2	a member of the VAV guanine nucleotide exchange factor family of oncogenes	0.312
cg14443041	Intergenic region		0.384

chromosomes. All these genes have been proposed as prognosis signature for the metastasis of breast cancer (Weigelt et al., 2005; Ma et al., 2003; van't Veer et al., 2002). Our results suggest the interplay of their expression with detected methylation sites. We list the functional annotation of probes detected on Chromosome 9 in Table 4 †.

In this analysis, we assume there is one pair of canonical directions between methylation and gene expression. We note that when the underlying canonical correlation structure is low-rank, the pair of canonical directions obtained from the proposed method lie in the subspace of true canonical directions. The extracted canonical directions can still be used to identify sets of methylate sites that are correlated with gene expression.

7. Proof of Main Theorem

We provide the proof of Theorem 4.1 in this section, which is based on the construction of an oracle sequence. The proof is similar in nature to that in Ma (2013) which focuses on the sparse PCA setting. Specifically, we are going to first define the strong signal sets, and then define an oracle sequence $(\alpha^{(k),ora}, \beta^{(k),ora})$ produced by Algorithms 1 and 2 only operating on the strong signal sets. We then show the desired rate of convergence for this oracle sequence. In the end, a probabilistic argument shows that with the help of thresholding, the oracle sequence is identical to the data-driven sequence with high probability. In the following proof, we condition on the second half of the data $(X_{n+1}, Y_{n+1}), ..., (X_{2n}, Y_{2n})$ and the event $\{||\hat{\Omega}_1 \Sigma_1 - I|| \vee ||\hat{\Omega}_2 \Sigma_2 - I|| = o(1)\}$. The "with probability" argument is understood to be with conditional probability unless otherwise specified. We keep using the notations $p = p_1 \vee p_2$ and $s = s_1 \vee s_2$.

7.1. Construction of the Oracle Sequence

We first define the strong signal set by

$$H_1 = \left\{ k : |\alpha_k| \geq \delta_1 \sqrt{\frac{\log p_1}{n}} \right\}, \quad H_2 = \left\{ k : |\beta_k| \geq \delta_2 \sqrt{\frac{\log p_2}{n}} \right\}. \tag{14}$$

We denote their complement in $\{1, 2, ..., p_1\}$ and $\{1, 2, ..., p_2\}$ by L_1 and L_2 respectively. Then we define the oracle version of $\hat{A}$ by taking those coordinates with strong signals. That is,

$$\hat{A}^{ora} = \begin{pmatrix} \hat{A}_{H_1 H_2} & 0 \\ 0 & 0 \end{pmatrix}.$$

---

† The corresponding canonical vector on genes RGS6, ORC6L, PTPRH, GPX2, QSOX2, NPM2, SCG3, RAB6B, L1CAM, STC2, REG1A is $(-0.252, -0.27, -0.286, -0.244, -0.35, -0.282, -0.256, -0.367, -0.256, 0.357, -0.358)$.

We construct the oracle initializer $(\alpha^{(0),ora}, \beta^{(0),ora})$ based on an oracle version of Algorithm 2 with the sets B_1 and B_2 replaced by $B_1^{ora} = B_1 \cap H_1$ and $B_2^{ora} = B_2 \cap H_2$. It is clear that $\alpha_{L_1}^{(0),ora} = 0$ and $\beta_{L_2}^{(0),ora} = 0$. Feeding the oracle initializer $(\alpha^{(0),ora}, \beta^{(0),ora})$ and the matrix $\hat{A}^{ora}$ into Algorithm 1, we get the oracle sequence $(\alpha^{(k),ora}, \beta^{(k),ora})$.

7.2. Data-Driven Thresholding

Algorithms 1 and 2 contain thresholding levels γ_1, γ_2 and t_{ij}. These tuning parameters can be specified by users. However, our theory is based on fully data-driven tuning parameters depending on the matrix $\hat{\Omega}_1 = (\hat{\omega}_{1,ij})$ and $\hat{\Omega}_2 = (\hat{\omega}_{2,ij})$. In particular, we use

$$t_{ij} = \frac{20\sqrt{2}}{9}\left(\sqrt{||\hat{\Omega}_1||\hat{\omega}_{2,jj}} + \sqrt{||\hat{\Omega}_2||\hat{\omega}_{1,ii}} + \sqrt{\hat{\omega}_{1,ii}\hat{\omega}_{2,jj}} + \sqrt{8||\hat{\Omega}_1||||\hat{\Omega}_2||/3}\right),$$

and

$$\gamma_1 = \left(0.17\min_{i,j} t_{ij}||\hat{\Omega}_2||^{1/2} + 2.1||\hat{\Omega}_2||^{1/2}||\hat{\Omega}_1||^{1/2} + 7.5||\hat{\Omega}_2||\right)\sqrt{\frac{\log p}{n}},$$

$$\gamma_2 = \left(0.17\min_{i,j} t_{ij}||\hat{\Omega}_1||^{1/2} + 2.1||\hat{\Omega}_1||^{1/2}||\hat{\Omega}_2||^{1/2} + 7.5||\hat{\Omega}_1||\right)\sqrt{\frac{\log p}{n}}.$$

The constants (δ_1, δ_2) in (14) are set as $\delta_1 = \delta_2 = 0.08w^{1/2}\min_{i,j} t_{ij}$. Such choice of thresholding levels are used in both the estimating sequence $(\alpha^{(k)}, \beta^{(k)})$ and the oracle sequence $(\alpha^{(k),ora}, \beta^{(k),ora})$.

7.3. Outline of Proof

The proof of Theorem 4.1 can be divided into the following three steps.

(a) Show that $\hat{A}^{ora}$ is a good approximation of $A = \lambda\alpha\beta^T$ in the sense that their first pairs of singular vectors are close. Namely, let $(\hat{\alpha}^{ora}, \hat{\beta}^{ora})$ be the first pair of singular vectors of $\hat{A}^{ora}$. We are going to bound $L(\hat{\alpha}^{ora}, \alpha)$ and $L(\hat{\beta}^{ora}, \beta)$.

(b) Show that the oracle sequence $(\alpha^{(k),ora}, \beta^{(k),ora})$ converges to $(\hat{\alpha}^{ora}, \hat{\beta}^{ora})$ after finite steps of iterations.

(c) Show that the estimating sequence $(\alpha^{(k)}, \beta^{(k)})$ and the oracle sequence $(\alpha^{(k),ora}, \beta^{(k),ora})$ are identical with high probability up to the necessary number of steps for convergence. Here, we need to first show that the oracle initializer $(\alpha^{(0),ora}, \beta^{(0),ora})$ is identical to the actual $(\alpha^{(0)}, \beta^{(0)})$. Then we are going to show the thresholding step in Algorithm 1 kills all the small coordinates so that the oracle sequence is identical to the estimating sequence under iteration.

7.4. Preparatory Lemmas

In this part, we present lemmas corresponding to the three steps in the outline of proof. The first lemma corresponds to Step 1.

LEMMA 7.1. *Under Assumptions A and B, we have*

$$
L(\hat{\alpha}^{ora}, \alpha)^2 \vee L(\hat{\beta}^{ora}, \beta)^2 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + ||\theta - \alpha|^2| \vee ||\eta - \beta||^2 \right),
$$

with probability at least $1 - O(p^{-2})$ *for some constant* $C > 0$.

Let $(\hat{l}_1, \hat{l}_2)$ be the first and second singular values of $\hat{A}^{ora}$. Then we have the following results, corresponding to Step 2.

LEMMA 7.2. *Under Assumptions A and B, we have*

$$
L(\alpha^{(1),ora}, \hat{\alpha}^{ora})^2 \leq 4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{32\gamma_1^2|H_1|}{|\hat{l}_1|^2},
$$

$$
L(\beta^{(1),ora}, \hat{\beta}^{ora})^2 \leq 4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{32\gamma_2^2|H_2|}{|\hat{l}_1|^2},
$$

for $k = 1$, *and*

$$
L(\alpha^{(k),ora}, \hat{\alpha}^{ora})^2 \leq \max\left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 \frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2} + \frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2}, \left(32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4\right)^{[k/2]} \right),
$$

$$
L(\beta^{(k),ora}, \hat{\beta}^{ora})^2 \leq \max\left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 \frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2} + \frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2}, \left(32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4\right)^{[k/2]} \right),
$$

for all $k \geq 2$ *with probability at least* $1 - O(p^{-2})$.

The quantities $|\hat{l}_1|$, $|\hat{l}_2/\hat{l}_1|$, $|H_1|$ and $|H_2|$ are determined by the following lemma.

LEMMA 7.3. *With probability at least* $1 - O(p^{-2})$,

$$
\begin{aligned}
|\hat{l}_2|^2 &\leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + ||\theta - \alpha||^2 \vee ||\eta - \beta||^2 \right), \\
|\hat{l}_1|^{-2} &\leq C.
\end{aligned}
$$

Moreover,

$$
|H_1| \leq C\left(s_1\left(\frac{\log p_1}{n}\right)^{-q/2} + \left(\frac{\log p_1}{n}\right)^{-1}||\theta - \alpha||^2 \right),
$$

$$
|H_2| \leq C\left(s_2\left(\frac{\log p_2}{n}\right)^{-q/2} + \left(\frac{\log p_2}{n}\right)^{-1}||\eta - \beta||^2 \right),
$$

for some constant $C > 0$.

Finally, we show that the oracle sequence and the actual sequence are identical with high probability, corresponding to Step 3. For the initializer, we have the following lemma.

LEMMA 7.4. *Under Assumptions A and B, we have $B_1^{ora} = B_1$ and $B_2^{ora} = B_2$ with probability at least $1 - O(p^{-2})$. Thus, $(\alpha^{(0),ora}, \beta^{(0),ora}) = (\alpha^{(0)}, \beta^{(0)})$.*

We proceed to analyze the sequence for $k \geq 1$ using mathematical induction. By iteration in Algorithm 1, we have

$$\alpha^{(k),ora} = \frac{T(\hat{A}^{ora}\beta^{(k-1),ora}, \gamma_1)}{||T(\hat{A}^{ora}\beta^{(k-1),ora}, \gamma_1)||}, \quad \beta^{(k),ora} = \frac{T(\hat{A}^{ora,T}\alpha^{(k-1),ora}, \gamma_2)}{||T(\hat{A}^{ora,T}\alpha^{(k-1),ora}, \gamma_2)||}.$$

Suppose we have $(\alpha^{(k-1),ora}, \beta^{(k-1),ora}) = (\alpha^{(k-1)}, \beta^{(k-1)})$. Then as long as

$$T(\hat{A}^{ora}\beta^{(k-1),ora}, \gamma_1) = T(\hat{A}\beta^{(k-1),ora}, \gamma_1) \tag{15}$$

$$T(\hat{A}^{ora,T}\alpha^{(k-1),ora}, \gamma_2) = T(\hat{A}^T\alpha^{(k-1),ora}, \gamma_2), \tag{16}$$

we have $(\alpha^{(k),ora}, \beta^{(k),ora}) = (\alpha^{(k)}, \beta^{(k)})$. Hence, it is sufficient to prove (15) and (16). Then, the result follows from mathematical induction. Without loss of generality, we analyze (15) as follows. Since $\beta_{L_2}^{(0),ora} = 0$, we may assume $\beta_{L_2}^{(k-1),ora} = 0$ at the k-th step. The vectors $\hat{A}^{ora}\beta^{(k-1),ora}$ and $\hat{A}\beta^{(k-1),ora}$ are respectively

$$\hat{A}^{ora}\beta^{(k-1),ora} = \begin{pmatrix} \hat{A}_{H_1 H_2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_{H_2}^{(k-1),ora} \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{A}_{H_1 H_2}\beta_{H_2}^{(k-1),ora} \\ 0 \end{pmatrix},$$

$$\hat{A}\beta^{(k-1),ora} = \begin{pmatrix} \hat{A}_{H_1 H_2} & \hat{A}_{H_1 L_2} \\ \hat{A}_{L_1 H_2} & \hat{A}_{L_1 L_2} \end{pmatrix} \begin{pmatrix} \beta_{H_2}^{(k-1),ora} \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{A}_{H_1 H_2}\beta_{H_2}^{(k-1),ora} \\ \hat{A}_{L_1 H_2}\beta_{H_2}^{(k-1),ora} \end{pmatrix}.$$

Hence, as long as $||\hat{A}_{L_1 H_2}\beta_{H_2}^{(k-1),ora}||_\infty \leq \gamma_1$, (15) holds. Similarly, as long as $||\hat{A}_{H_1 L_2}^T\alpha_{H_1}^{(k-1),ora}||_\infty \leq \gamma_2$, (16) holds. This is guaranteed by the following lemma.

LEMMA 7.5. *For any sequence of unit vectors $(a^{(k)}, b^{(k)}) \in \mathbb{R}^{|H_1|} \times \mathbb{R}^{|H_2|}$, with $k = 1, 2, ..., K$ for some $K = O(1)$. We assume that they only depend on $\hat{A}_{H_1 H_2}$. Then, under the current choice of (γ_1, γ_2), we have*

$$||\hat{A}_{L_1 H_2}b^{(k)}||_\infty \leq \gamma_1, \quad ||\hat{A}_{H_1 L_2}^T a^{(k)}||_\infty \leq \gamma_2,$$

for all $k = 1, ..., K$ with probability at least $1 - O(p^{-2})$.

7.5. *Proof of Theorem 4.1*

PROOF (PROOF OF THEOREM 4.1). In the following proof, C denotes a generic constant which may vary from line to line. Without loss of generality, we prove convergence of

$\alpha^{(k)}$. By Lemma 7.4 and Lemma 7.5, $\alpha^{(k)} = \alpha^{(k),ora}$ for all $k = 1, 2, ..., K$ with probability $1 - O(p^{-2})$. Hence, it is sufficient to prove convergence of $\alpha^{(k),ora}$. Lemma 7.2 implies that for $k = 2, 3, ..., K$,

$$
\begin{aligned}
L(\alpha^{(k),ora}, \hat{\alpha}^{ora})^2 &\leq \max\left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 \frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2} + \frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2}, \left(32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4\right)^{[k/2]} \right) \\
&\leq \max\left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 \frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2} + \frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2}, 32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4 \right)
\end{aligned}
$$

According to Lemma 7.3, we have

$$
32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + ||\theta - \alpha||^2 \vee ||\beta - \eta||^2 \right)^2.
$$

We also have

$$
\begin{aligned}
&4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 \frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2} + \frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2} \\
&\leq C\left(\gamma_2^2|H_2| + \gamma_1^2|H_1| \right) \\
&\leq C\frac{\log p}{n}\left(s\left(\frac{\log p}{n}\right)^{-q/2} + \left(\frac{\log p}{n}\right)^{-1}||\theta - \alpha||^2 \right) \\
&\leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + ||\theta - \alpha||^2 \right).
\end{aligned}
$$

Hence, the desired bound holds for $k = 2, 3, ..., K$. For $k = 1$, by Lemma 7.2, we have

$$
L(\alpha^{(1),ora}, \hat{\alpha}^{ora})^2 \leq 4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{32\gamma_1^2|H_1|}{|\hat{l}_1|^2} \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + ||\theta - \alpha||^2 \right).
$$

Therefore, we have proved the bound of $L(\alpha^{(k),ora}, \alpha^{ora})^2$ for all $k = 1, 2, ..., K$. Combining this result and Lemma 7.1, we have

$$
L(\alpha^{(k),ora}, \alpha)^2 \leq C\left(s\left(\frac{\log p}{n}\right)^{1-q/2} + ||\theta - \alpha||^2 \vee ||\beta - \eta||^2 \right),
$$

for $k = 1, 2, ..., K$. The final bound follows from the triangle inequality applied to the equation above and the fact that

$$
L(\alpha, \theta) \leq C\left(||\alpha - \theta|| \wedge ||\alpha + \theta|| \right) \leq C||\alpha - \theta||.
$$

The same analysis applies for $L(\beta^{(k),ora}, \eta)^2$. Thus, the result is obtained conditioning on

$$
(X_{n+1}, Y_{n+1}), ..., (X_{2n}, Y_{2n}).
$$

Since we assume $\xi_\Omega = o(1)$ with probability at least $1 - O(p^{-2})$, the unconditional result also holds.

Appendix

A.   Technical Lemmas

We define the high-signal and low-signal set of (θ, η) by

$$H_1' = \left\{ k : |\theta_k| \geq \frac{1}{2} \delta_1 \sqrt{\frac{\log p_1}{n}} \right\}, \quad H_2' = \left\{ k : |\eta_k| \geq \frac{1}{2} \delta_2 \sqrt{\frac{\log p_2}{n}} \right\},$$

and $L_1' = \{1, ..., p_1\} - H_1'$ and $L_2' = \{1, ..., p_2\} - H_2'$.

LEMMA A.1. *We have*

$$|H_1'| \leq (\delta_1/2)^{-q} s_1 \left(\frac{\log p_1}{n} \right)^{-q/2}, \quad |H_2'| \leq (\delta_2/2)^{-q} s_2 \left(\frac{\log p_2}{n} \right)^{-q/2},$$

$$|H_1| \leq (\delta_1/2)^{-q} s_1 \left(\frac{\log p_1}{n} \right)^{-q/2} + (\delta_1/2)^{-2} \left(\frac{\log p_1}{n} \right)^{-1} ||\theta - \alpha||^2,$$

$$|H_2| \leq (\delta_2/2)^{-q} s_2 \left(\frac{\log p_2}{n} \right)^{-q/2} + (\delta_2/2)^{-2} \left(\frac{\log p_2}{n} \right)^{-1} ||\eta - \beta||^2,$$

$$|L_1 - L_1'| = |H_1' - H_1| \leq 2(\delta_1/2)^{-q} s_1 \left(\frac{\log p_1}{n} \right)^{-q/2} + (\delta_1/2)^{-2} \left(\frac{\log p_1}{n} \right)^{-1} ||\theta - \alpha||^2,$$

$$|L_2 - L_2'| = |H_2' - H_2| \leq 2(\delta_2/2)^{-q} s_2 \left(\frac{\log p_2}{n} \right)^{-q/2} + (\delta_2/2)^{-2} \left(\frac{\log p_2}{n} \right)^{-1} ||\eta - \beta||^2.$$

For the transformed data $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^n$, it has a latent variable representation

$$\tilde{X}_i = \sqrt{\lambda} \alpha Z_i + X_i', \quad \tilde{Y}_i = \sqrt{\lambda} \beta Z_i + Y_i',$$

where Z_i, X_i', Y_i' are independent, $Z_i \sim N(0, 1)$ and X_i' and Y_i' are Gaussian vectors.

LEMMA A.2. *The latent representation above exists in the sense that* $\mathrm{Cov}(X') \geq 0$ *and* $\mathrm{Cov}(Y') \geq 0$. *Moreover, we have*

$$||\mathrm{Cov}(X')|| \leq (1 + o(1)) ||\hat{\Omega}_1||, \quad \text{and} \quad ||\mathrm{Cov}(Y')|| \leq (1 + o(1)) ||\hat{\Omega}_2||.$$

LEMMA A.3 (JOHNSTONE (2001)). *Given* $Z_1, ..., Z_n$ *i.i.d.* $N(0, 1)$. *For each* $t \in (0, 1/2)$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n Z_i^2 - 1 \right| > t \right) \leq 2 \exp\left(-\frac{3nt^2}{16} \right).$$

LEMMA A.4. *Let* X_{H_1}' *be an* $n \times |H_1|$ *matrix with* X_{i,H_1}' *being the i-th row and* Y_{H_2}' *be an* $n \times |H_2|$ *matrix with* Y_{i,H_2}' *being the i-th row. We have for any* $t > 0$,

$$\mathbb{P}\left(||X_{H_1}'^T X_{H_1}'|| > 1.01 ||\hat{\Omega}_1|| \left(n + 2(\sqrt{n|H_1|} + nt) + (\sqrt{|H_1|} + \sqrt{n}t)^2 \right) \right) \leq 2e^{-nt^2/2},$$

$$\mathbb{P}\left(||Y_{H_2}'^{T}Y_{H_2}'|| > 1.01||\hat{\Omega}_2||\left(n + 2\left(\sqrt{n|H_2|} + nt\right) + \left(\sqrt{|H_2|} + \sqrt{n}t\right)^2\right)\right) \le 2e^{-nt^2/2},$$

$$\mathbb{P}\left(||X_{H_1}'^{T}Y_{H_2}'|| > 1.03||\hat{\Omega}_1||^{1/2}||\hat{\Omega}_2||^{1/2}\left(\sqrt{|H_1|n} + \sqrt{|H_2|n} + t\sqrt{n}\right)\right) \le (|H_1|\wedge|H_2|)e^{-3n/64} + e^{-t^2/2}.$$

LEMMA A.5. *We have for any* $t > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} Z_i X_{i,H_1}'\right\| > 1.03||\hat{\Omega}_1||^{1/2}\left((t+1)\sqrt{n} + \sqrt{|H_1|n}\right)\right) \le e^{-3n/64} + e^{-t^2/2},$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} Z_i Y_{i,H_2}'\right\| > 1.03||\hat{\Omega}_2||^{1/2}\left((t+1)\sqrt{n} + \sqrt{|H_2|n}\right)\right) \le e^{-3n/64} + e^{-t^2/2},$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} Z_i X_{i,L_1}'\right\|_{\infty} > 1.03||\hat{\Omega}_1||^{1/2}(t+2)\sqrt{n}\right) \le |L_1|\left(e^{-3n/64} + e^{-t^2/2}\right),$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} Z_i Y_{i,L_2}'\right\|_{\infty} > 1.03||\hat{\Omega}_2||^{1/2}(t+2)\sqrt{n}\right) \le |L_2|\left(e^{-3n/64} + e^{-t^2/2}\right).$$

LEMMA A.6. *We have*

$$||\hat{A}_{H_1 H_2} - A_{H_1 H_2}|| \le C\left(s^{1/2}\left(\frac{\log p}{n}\right)^{1/2-q/4} + ||\alpha - \theta|| \vee ||\beta - \eta||\right),$$

$$\max_{i=1,2}|\hat{l}_i - l_i| \le C\left(s^{1/2}\left(\frac{\log p}{n}\right)^{1/2-q/4} + ||\alpha - \theta|| \vee ||\beta - \eta||\right),$$

$$l_1 \ge C^{-1},$$

$$l_2 = 0$$

for some constant $C > 0$ *with probability at least* $1 - O(p^{-2})$. *The quantities* $\hat{l}_i$ *and* l_i *are the i-th singular values of* $\hat{A}_{H_1 H_2}$ *and* $A_{H_1 H_2}$ *respectively.*

B. Analysis of the Initializer

We show that the Algorithm 2 actually returns a consistent estimator of leading pair of singular vectors $\left(\alpha^{(0)}, \beta^{(0)}\right)$, which serves as a good candidate for the power method in Algorithm 1. To be specific, with high probability, the initialization procedure correctly kills all low signal coordinates to zeros, i.e. $\alpha_{L_1}^{(0)} = 0$ and $\beta_{L_2}^{(0)} = 0$, by the thresholding step. On the other hand, although Algorithm 2 cannot always correctly pick all strong signal coordinates, it does pick those much stronger ones such that $\left(\alpha^{(0)}, \beta^{(0)}\right)$ is still consistent up to a sign. The properties are summarized below.

LEMMA B.1. *With probability* $1 - Cp^{-2}$, *we have that,*

(a) $B_i \subset H_i$, $i = 1, 2$;

(b) $\left| \hat{l}_i - \hat{l}_i^B \right| \to 0$ for $i = 1, 2$. where $\hat{l}_i$ and $\hat{l}_i^B$ are the ith singular value of $\hat{A}_{H_1 H_2}$ and $\hat{A}_{B_1 B_2}$;

(c) $\alpha^{(0)}$ and $\beta^{(0)}$ are consistent, i.e., $L\left(\alpha^{(0)}, \hat{\alpha}^{ora}\right) \to 0$ and $L\left(\beta^{(0)}, \hat{\beta}^{ora}\right) \to 0$.

The procedure is to select those strong signal coordinates of α and β and is similar to the "diagonal thresholding" sparse PCA method proposed by Johnstone and Lu (2009). However, unlike the PCA setting, we cannot get the information of each coordinate through its corresponding diagonal entry of the sample covariance matrix. Instead, we measure the strength of coordinates in terms of the maximum entry among its corresponding row or column of the sample covariance matrix and still can capture all coordinates of α and β above the level of $\left(\frac{\log p}{n}\right)^{1/4}$. The sparsity assumption in Equation (10) is needed to guarantee the consistency of the initial estimator $\left(\alpha^{(0)}, \beta^{(0)}\right)$.

The proof is developed in the following. In the first part, we prove the three results in Lemma B.1 along with stating some useful propositions. We then prove those propositions in the second part.

B.1.   Proof of Lemma B.1

PROOF (PROOF OF RESULT 1). We start with showing the first result $B_i \subset H_i$. In fact we show that the index set B_i screens out all weak signal coordinates as well as captures all coordinates with much stronger signal coordinates $\sqrt{\frac{\log p}{n}} s^{\frac{1}{2-q}}$, compared with the thresholding $\sqrt{\frac{\log p}{n}}$ of H_i. By the sparsity assumption (10), clearly B_i also captures all coordinates with much stronger signal $\left(\frac{\log p}{n}\right)^{1/4}$. To be specific, we show that with probability $1 - Cp^{-2}$, we have

$$B_i^- \subset B_i \subset H_i, \tag{17}$$

where the index set of those stronger signal coordinates are defined as follows,

$$B_1^- = \left\{ i, |\alpha_i| > \phi\sqrt{\frac{\log p}{n}} s_2^{\frac{1}{2-q}} \right\}, B_2^- = \left\{ i, |\beta_i| > \phi\sqrt{\frac{\log p}{n}} s_1^{\frac{1}{2-q}} \right\}.$$

The constant ϕ is determined in the analysis. $\phi = 2\max_{i,j} t_{ij} C_\lambda^{-1} \left(\frac{\sqrt{W}}{0.6} \left(2W^{1/2}\right)^{\frac{-q}{2-q}} \right)^{1/2}$.

Recall that $\hat{A} = \frac{1}{n} \sum_i \tilde{X}_i \tilde{Y}_i^T$ with conditional mean $\mathbb{E}\hat{A} = A = \lambda\alpha\beta^T = (a_{ij})_{p_1 \times p_2}$. The result (17) can be shown in two steps. During the first step we show that with probability $1 - Cp^{-2}$, the index set B_i satisfies $B_i^{--} \subset B_i \subset B_i^{++}$, where for simplicity we pick $\kappa_- = 0.09, \kappa_+ = 2$ and

$$B_1^{++} = \left\{ i : \max_j \frac{|a_{ij}|}{t_{ij}} > \kappa_- \sqrt{\frac{1}{n}\log p} \right\}, B_1^{--} = \left\{ i : \max_j \frac{|a_{ij}|}{t_{ij}} > \kappa_+ \sqrt{\frac{1}{n}\log p} \right\},$$

$$B_2^{++} = \left\{ j : \max_i \frac{|a_{ij}|}{t_{ij}} > \kappa_- \sqrt{\frac{1}{n} \log p} \right\}, B_2^{--} = \left\{ j : \max_i \frac{|a_{ij}|}{t_{ij}} > \kappa_+ \sqrt{\frac{1}{n} \log p} \right\}.$$

For the second step, we show that $B_i^{++} \subset H_i$ and $B_i^{--} \subset B_i^-$ with probability $1 - Cp^{-2}$. We present these two results in the following two propositions.

PROPOSITION B.1. *With probability* $1 - Cp^{-2}$*, we have* $B_i^{--} \subset B_i \subset B_i^{++}$ *for* $i = 1, 2$.

PROPOSITION B.2. *With probability* $1 - Cp^{-2}$*, we have* $B_i^{++} \subset H_i$ *and* $B_i^{--} \subset B_i^-$ *for* $i = 1, 2$.

Thus, the proof is complete.

Before proving Result 2 and Result 3, we need the following proposition.

PROPOSITION B.3. *Define*

$$e_B^2 = \max \left\{ \sum_{i \in \left(B_1^-\right)^c} \alpha_i^2, \sum_{i \in \left(B_2^-\right)^c} \beta_i^2 \right\}. \tag{18}$$

Then we have for some constants $C_1, C_2 > 0$*,*

$$e_B^2 \le C_1 \left(\frac{\log p}{n} \right)^{1-q/2} \left(s_1^2 \vee s_2^2 \right) + C_2 \xi_\Omega^2. \tag{19}$$

Moreover, under our assumptions (10) and (11), we obtain $e_B = o(1)$ *with probability* $1 - Cp^{-2}$*.*

Now we restrict our attention on the event on which the result (17) holds, which is valid with high probability $1 - Cp^{-2}$. Define index set $D_i = H_i \backslash B_i$ and

$$\hat{A}_{H_1 H_2}^{B_1 B_2} = \begin{pmatrix} \hat{A}_{B_1 B_2} & 0 \\ 0 & 0 \end{pmatrix}_{|H_1| \times |H_2|}.$$

PROOF (PROOF OF RESULT 2). We show the second result $\left| \hat{l}_i - \hat{l}_i^B \right| \to 0$ for $i = 1, 2$. Note that $\hat{l}_i^B$ is the ith singular value of $\hat{A}_{B_1 B_2}$ and hence is also the ith singular value of $\hat{A}_{H_1 H_2}^{B_1 B_2}$. Applying Weyl's theorem, we obtain that

$$\left| \hat{l}_i - \hat{l}_i^B \right| \le \left\| \hat{A}_{H_1 H_2} - \hat{A}_{H_1 H_2}^{B_1 B_2} \right\| \le \left\| \hat{A}_{D_1 B_2} \right\| + \left\| \hat{A}_{B_1 D_2} \right\| + \left\| \hat{A}_{D_1 D_2} \right\|. \tag{20}$$

To bound $\left\| \hat{A}_{D_1 B_2} \right\|$, we apply the latent variable representation in Lemma A.2 and obtain that $\hat{A}_{D_1 B_2} = \sum_{j=1}^4 G_j$, where

$$G_1 = \frac{\lambda Z^T Z}{n} \alpha_{D_1} \beta_{B_2}^T, G_2 = \frac{1}{n} X_{D_1}'^T Y_{B_2}', G_3 = \frac{\sqrt{\lambda} \alpha_{D_1} Z^T Y_{B_2}'}{n}, G_4 = \frac{\sqrt{\lambda} X_{D_1}'^T Z \beta_{B_2}^T}{n}.$$

Now we bound G_j separately as follows. According to Lemma A.3, Proposition B.3 and the fact $\|\beta_{B_2}\| \leq \|\beta\| \leq \left\|\hat{\Omega}_2\Sigma_2\right\| \|\eta\| = (1+o(1))\|\eta\|$, we obtain that with probability $1 - Cp^{-2}$

$$\begin{aligned}
\|G_1\| &\leq \lambda\left(1 + O(\sqrt{\frac{\log p}{n}})\right)\|\alpha_{D_1}\| \|\beta_{B_2}\| \leq (1+o(1))\lambda\|\eta\| \left\|\alpha_{(B_1^-)^c}\right\| \\
&\leq (1+o(1))\lambda\|\theta\| e_B,
\end{aligned}$$

where the second inequality follows from $D_1 \subset (B_1^-)^c$. The third inequality in Lemma A.4 implies that with probability $1 - Cp^{-2}$,

$$\|G_2\| \leq \left(C(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{|H_1|}{n}} + \sqrt{\frac{|H_2|}{n}})\right) = o(e_B),$$

where the last inequality is due to Lemma A.1 $|H_i| = O\left(s_i\left(\frac{\log p}{n}\right)^{-q/2}\right)$ and Proposition B.3. Moreover, the first two inequalities in Lemma A.5 further imply that with probability $1 - Cp^{-2}$,

$$\begin{aligned}
\|G_3\| &\leq \left(C(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{|H_2|}{n}})\|\alpha_{D_1}\|\right) \leq o(\left\|\alpha_{(B_1^-)^c}\right\|) = o(e_B), \\
\|G_4\| &\leq \left(C(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{|H_1|}{n}})\|\beta_{B_2}\|\right) = o(e_B).
\end{aligned}$$

Combining the above four results, we obtain that $\left\|\hat{A}_{D_1 B_2}\right\| \leq (1+o(1))\lambda\|\theta\| e_B$ with probability $1 - Cp^{-2}$. Similarly we can obtain that $\left\|\hat{A}_{B_1 D_2}\right\| \leq (1+o(1))\lambda\|\theta\| e_B$ with probability $1 - Cp^{-2}$. To bound $\left\|\hat{A}_{D_1 D_2}\right\|$, similarly we can write $\hat{A}_{D_1 B_2} = \sum_{j=1}^{4} F_j$, where

$$F_1 = \frac{\lambda Z^T Z}{n}\alpha_{D_1}\beta_{D_2}^T, F_2 = \frac{1}{n}X_{D_1}'^T Y_{D_2}', F_3 = \frac{\sqrt{\lambda}\alpha_{D_1} Z^T Y_{D_2}'}{n}, F_4 = \frac{\sqrt{\lambda}X_{D_1}'^T Z\beta_{D_2}^T}{n}.$$

Note that $D_i \subset (B_i^-)^c$ for $i = 1, 2$. Lemma A.4, Lemma A.5 and Proposition B.3 imply that with probability $1 - Cp^{-2}$,

$$\begin{aligned}
\|F_1\| &\leq \lambda\left(1 + O(\sqrt{\frac{\log p}{n}})\right)\|\alpha_{D_1}\|\|\beta_{D_2}\| \leq C\left\|\alpha_{(B_1^-)^c}\right\|\left\|\beta_{(B_2^-)^c}\right\| = o(e_B), \\
\|F_2\| &\leq \left(C(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{|H_1|}{n}} + \sqrt{\frac{|H_2|}{n}})\right) = o(e_B), \\
\|F_3\| &\leq o(\left\|\alpha_{(B_1^-)^c}\right\|) = o(e_B), \|F_4\| \leq o(\left\|\beta_{(B_2^-)^c}\right\|) = o(e_B).
\end{aligned}$$

Hence we obtain $\left\|\hat{A}_{D_1 D_2}\right\| = o(e_B)$. These upper bounds for $\left\|\hat{A}_{D_1 D_2}\right\|$, $\left\|\hat{A}_{D_1 D_2}\right\|$ and $\left\|\hat{A}_{D_1 D_2}\right\|$, together with Equation (20) imply that with probability $1 - Cp^{-2}$,

$$\left|\hat{l}_i - \hat{l}_i^B\right| \leq \left\|\hat{A}_{H_1 H_2} - \hat{A}_{H_1 H_2}^{B_1 B_2}\right\| \leq (1+o(1))(\|\theta\| + \|\eta\|)\lambda e_B = o(1), \tag{21}$$

where the last inequality follows from Proposition B.3 and Assumption B ($\max \{\|\theta\|, \|\eta\|\} \leq w^{-1/2}$).

PROOF (PROOF OF RESULT 3). We show that last result $L\left(\alpha^{(0)}, \hat{\alpha}^{ora}\right) \to 0$ and $L\left(\beta^{(0)}, \hat{\beta}^{ora}\right) \to 0$. Note $\alpha^{(0)}$ and $\hat{\alpha}^{ora} \in \mathbb{R}^{p_1}$ but all entries in the index set H_1^c are zeros. Hence we only need to compare them in $\mathbb{R}^{|H_1|}$. Similarly we calculate $L\left(\beta^{(0)}, \hat{\beta}^{ora}\right)$ in space $\mathbb{R}^{|H_2|}$. Constraint on coordinates in $H_1 \times H_2$, $\left(\alpha^{(0)}, \beta^{(0)}\right)$ and $\left(\hat{\alpha}^{ora}, \hat{\beta}^{ora}\right)$ are leading pair of singular vectors of $\hat{A}_{H_1 H_2}^{B_1 B_2}$ and $\hat{A}_{H_1 H_2}$ respectively. We apply Wedin's theorem (See Stewart and Sun (1990), Theorem 4.4) to $\hat{A}_{H_1 H_2}$ and $\hat{A}_{H_1 H_2}^{B_1 B_2}$ to obtain that

$$
\begin{aligned}
\max\left\{L\left(\alpha^{(0)}, \hat{\alpha}^{ora}\right), L\left(\beta^{(0)}, \hat{\beta}^{ora}\right)\right\} &\leq \sqrt{2} \max\left\{\left\|\frac{\alpha^{(0)}\left(\alpha^{(0)}\right)^T}{\|\alpha^{(0)}\|^2} - \frac{\hat{\alpha}^{ora}\left(\hat{\alpha}^{ora}\right)^T}{\|\hat{\alpha}^{ora}\|^2}\right\|,\right.\\
&\qquad\qquad \left.\left\|\frac{\beta^{(0)}\left(\beta^{(0)}\right)^T}{\|\beta^{(0)}\|^2} - \frac{\hat{\beta}^{ora}\left(\hat{\beta}^{ora}\right)^T}{\|\hat{\beta}^{ora}\|^2}\right\|\right\}\\
&\leq \sqrt{2}\frac{\left\|\hat{A}_{H_1 H_2}^{B_1 B_2} - \hat{A}_{H_1 H_2}\right\|}{\delta\left(\hat{A}_{H_1 H_2}^{B_1 B_2}, \hat{A}_{H_1 H_2}\right)}, \qquad (22)
\end{aligned}
$$

where $\delta\left(\hat{A}_{H_1 H_2}^{B_1 B_2}, \hat{A}_{H_1 H_2}\right) = \hat{l}_1^B - \hat{l}_2$. The result of Lemma A.6 implies that $\hat{l}_2 = o(1)$, and Result 2 we just showed further implies that $\hat{l}_1^B = (1 + o(1))\hat{l}_1 = (1 + o(1))\lambda \|\eta\| \|\theta\|$ with probability $1 - Cp^{-2}$. Therefore we obtain that with probability $1 - Cp^{-2}$, the value $\delta\left(\hat{A}_{H_1 H_2}^{B_1 B_2}, \hat{A}_{H_1 H_2}\right) = (1 + o(1))\lambda \|\eta\| \|\theta\|$ is bounded below and above by constants according to Assumption B. This fact, together with Equations (21) and (22) completes our proof

$$
\max\left\{L\left(\alpha^{(0)}, \hat{\alpha}^{ora}\right), L\left(\beta^{(0)}, \hat{\beta}^{ora}\right)\right\} = o(1),
$$

with probability $1 - Cp^{-2}$.

B.2. Proofs of Propositions

PROOF (PROOF OF PROPOSITION B.1). We first provide concentration inequality for each $\hat{a}_{ij} = \frac{1}{n}\sum_k \tilde{X}_{k,i}\tilde{Y}_{k,j}$. By the latent variable representation, we have $\tilde{X}_{1,i} = \sqrt{\lambda}\alpha_i Z_1 + X'_{1,i}$ and $\tilde{Y}_{1,j} = \sqrt{\lambda}\beta_j Z_1 + Y'_{1,j}$ where $Z_1 \sim N(0,1)$, $X'_{1,i} \sim N\left(0, \mathrm{Var}\left(X'_{1,i}\right)\right)$ and $Y'_{1,j} \sim N\left(0, \mathrm{Var}\left(Y'_{1,j}\right)\right)$ are independent. This representation leads to

$$
\hat{a}_{ij} = \frac{1}{n}\left(\lambda\alpha_i\beta_j\sum_k Z_k^2 + \sum_k X'_{k,i}Y'_{k,j} + \sqrt{\lambda}\alpha_i\sum_k Z_k Y'_{k,j} + \sqrt{\lambda}\beta_j\sum_k Z_k X'_{k,i}\right).
$$

The consistency assumption Equation (11) in Assumption B implies that $\mathrm{Var}\left(X'_{1,i}\right) \leq$ $\mathrm{Var}\left(\tilde{X}_{1,i}\right) = (1 + o(1))\hat{\omega}_{1,ii} = (1 + o(1))\omega_{1,ii}$. Applying Lemma A.3, we obtain that

$$\mathbb{P}\left(\left|\frac{\lambda\alpha_i\beta_j}{n}\sum_k Z_k^2 - a_{ij}\right| > \lambda\alpha_i\beta_j t\right) \leq 2\exp\left(-\frac{3nt^2}{16}\right). \tag{23}$$

Following the line of proof in Lemma A.5 and Proposition D.2 in Ma (2013), for n and p large enough (hence $n^{-1}\log p \to 0$), we have the following concentration inequalities,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_k X'_{k,i}Y'_{k,j}\right| > (1 + o(1))\,b\sqrt{\hat{\omega}_{1,ii}\hat{\omega}_{2,jj}}\sqrt{\frac{\log p}{n}}\right) \quad\leq\quad 2p^{-b^2/2}, \tag{24}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sqrt{\lambda}\beta_j\sum_k Z_k X'_{k,i}\right| > (1 + o(1))\,b\sqrt{\lambda}\,|\beta_j|\,\sqrt{\hat{\omega}_{1,ii}}\sqrt{\frac{\log p}{n}}\right) \quad\leq\quad 2p^{-b^2/2}, \tag{25}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sqrt{\lambda}\alpha_i\sum_k Z_k Y'_{k,j}\right| > (1 + o(1))\,b\sqrt{\lambda}\,|\alpha_i|\,\sqrt{\hat{\omega}_{2,jj}}\sqrt{\frac{\log p}{n}}\right) \quad\leq\quad 2p^{-b^2/2}. \tag{26}$$

Recall the definition of the adaptive thresholding level t_{ij},

$$t_{ij} = \tfrac{20\sqrt{2}}{9}\left(\sqrt{\lambda_{\max}\left(\hat{\Omega}_1\right)\hat{\omega}_{2,jj}} + \sqrt{\lambda_{\max}\left(\hat{\Omega}_2\right)\hat{\omega}_{1,ii}}\right.$$
$$\left. + \sqrt{\hat{\omega}_{1,ii}\hat{\omega}_{2,jj}} + \sqrt{8\lambda_{\max}\left(\hat{\Omega}_1\right)\lambda_{\max}\left(\hat{\Omega}_2\right)/3}\right).$$

Applying the union bound to Equations (23)-(26), we obtain the concentration inequality for $\hat{a}_{ij}$ as follows

$$\mathbb{P}\left(\frac{|\hat{a}_{ij} - a_{ij}|}{t_{ij}} > (1 + o(1))\frac{9b_1}{20\sqrt{2}}\sqrt{\frac{\log p}{n}}\right) \leq 8p^{-b_1^2/2}, \tag{27}$$

where we used the fact $\lambda \leq 1$ and $|\alpha_i| \leq (1 + o(1))|\theta_i| \leq (1 + o(1))\lambda_{\max}^{1/2}\left(\hat{\Omega}_1\right)$.

We finish our proof by bounding the probability $\mathbb{P}\left(B_i \not\subset B_i^{++}\right)$ and $\mathbb{P}\left(B_i^{--} \not\subset B_i\right)$ respectively. Let j_i^* be an integer such that $|a_{ij_i^*}|/t_{ij_i^*} = \max_j |a_{ij}|/t_{ij}$. We apply the union bound to obtain

$$\begin{aligned}
\mathbb{P}\left(B_1^{--} \not\subset B_1\right) &\leq \sum_{i\in B_1^{--}} \mathbb{P}\left\{\max_j \frac{|\hat{a}_{ij}|}{t_{ij}} \leq \sqrt{\frac{\log p}{n}}\right\} \\
&\leq \sum_{i\in B_1^{--}} \mathbb{P}\left\{|\hat{a}_{ij_i^*}|/t_{ij_i^*} \leq \sqrt{\frac{\log p}{n}}\right\} \\
&\leq \sum_{i\in B_1^{--}} \mathbb{P}\left\{|a_{ij_i^*} - \hat{a}_{ij_i^*}|/t_{ij_i^*} > (\kappa_+ - 1)\sqrt{\frac{\log p}{n}}\right\} \\
&\leq p8p^{-b_1^2/2} \leq Cp^{-3},
\end{aligned}$$

where the last inequality follows from Equation (27) with $b_1 = 2\sqrt{2}$ and $\kappa_+ = 2$. Similarly,

we apply the union bound to obtain

$$
\begin{aligned}
\mathbb{P}\left(B_1 \not\subset B_1^{++}\right) &\leq \sum_{i \in \left(B_1^{++}\right)^c} \mathbb{P}\left\{\max_j \frac{|\hat{a}_{ij}|}{t_{ij}} > \sqrt{\frac{\log p}{n}}\right\} \\
&\leq \sum_{i \in \left(B_1^{++}\right)^c} \sum_{j=1}^{p_2} \mathbb{P}\left\{|\hat{a}_{ij}|/t_{ij} > \sqrt{\frac{\log p}{n}}\right\} \\
&\leq \sum_{i \in \left(B_1^{++}\right)^c} \sum_{j=1}^{p_2} \mathbb{P}\left\{|\hat{a}_{ij} - a_{ij}|/t_{ij} > (1 - \kappa_-)\sqrt{\frac{\log p}{n}}\right\} \\
&\leq p^2 8 p^{-b_1^2/2} \leq C p^{-2},
\end{aligned}
$$

where the last inequality follows from Equation (27) with $b_1 = 2\sqrt{2}$ and $\kappa_- = 0.09$. We can obtain the bounds for $\mathbb{P}\left(B_2 \not\subset B_2^{++}\right)$ and $\mathbb{P}\left(B_2^{--} \not\subset B_2\right)$, which finish our proof.

PROOF (PROOF OF PROPOSITION B.2). Recall $a_{ij} = \lambda \alpha_i \beta_j$. To show $B_1^{++} \subset H_1$, we only need to show that $\frac{\kappa_- \min_{i,j} t_{ij}}{\lambda \max_j |\beta_j|} > \delta_1$, noting $p_1 \leq p$. To see this, the key part is to bound $\max_j |\beta_j|$ from above. In fact, Assumption B implies $\max_j |\beta_j| \leq \|\beta\| \leq \left\|\hat{\Omega}_2 \Sigma_2\right\| \|\eta\| = (1 + o(1)) w^{-1/2}$ with probability $1 - C p^{-2}$. Therefore this upper bound of $\|\beta\|$, the definition of δ_1 and $\lambda \leq 1$ imply that,

$$
\frac{\kappa_- \min_j t_{ij}}{\lambda \max_j |\beta_j|} \geq 0.08 w^{1/2} \min_{i,j} t_{ij} = \delta_1,
$$

where the last equation follows from the definition of δ_1. Similarly, we can show that $B_2^{++} \subset H_2$ with probability $1 - C p^{-2}$.

To show $B_1^{--} \subset B_1^-$, we only need to show that $\phi_1 s_2^{\frac{1}{2-q}} > \frac{\kappa_+ \max_{i,j} t_{ij}}{\lambda \max_j |\beta_j|}$. This time the key part is to bound $\max_j |\beta_j|$ from below. Note that $\|\beta\| \geq (1 - o(1))\|\eta\| \geq (1 - o(1)) W^{-1/2}$ follows from Assumption B. For any positive integer k, we denote I_k as the index set of the largest k coordinates of η in magnitude. Then we have with probability $1 - C p^{-2}$,

$$
\begin{aligned}
0.9 W^{-1/2} &\leq \|\beta\| \leq k \max_j |\beta_j|^2 + \sum_{j \in I_k^c} |\beta_j|^2 \\
&\leq k \max_j |\beta_j|^2 + \sum_{j \in I_k^c} |\eta_j|^2 + \|\beta - \eta\| \\
&\leq k \max_j |\beta_j|^2 + \frac{q}{2-q} s_2^{2/q} k^{1-2/q} + o(1).
\end{aligned}
$$

Picking $k_0 = \left\lceil s_2^{\frac{2}{2-q}} \left(2 W^{1/2}\right)^{\frac{q}{2-q}} \right\rceil$, the Equation above implies that $k_0 \max_j |\beta_j|^2 \geq 0.3 W^{-1/2}$. Consequently, we get a lower bound $\max_j |\beta_j| \geq \sqrt{C_l} s_2^{-\frac{1}{2-q}}$, where constant $C_l = \frac{0.6}{\sqrt{W}} \left(2 W^{1/2}\right)^{\frac{q}{2-q}}$. We complete our proof by noting that the lower bound of $\max_j |\beta_j|$, the definition of ϕ and Assumption $\lambda > C_\lambda$ imply

$$
\frac{\kappa_+ \max_{i,j} t_{ij}}{\lambda \max_j |\beta_j|} \leq 2 \max_{i,j} t_{ij} C_\lambda^{-1} C_l^{-1/2} s_2^{\frac{1}{2-q}} = \phi s_2^{\frac{1}{2-q}},
$$

where the last equation follows from the definition of constant ϕ. Similarly, we can show $B_2^{--} \subset B_2^{-}$ with probability $1 - Cp^{-2}$.

PROOF (PROOF OF PROPOSITION B.3). Note we only assume θ and η are in the weak l_q ball. Define the relatively weak signal coordinates of θ and η as

$$B_1'^{-} = \left\{ i, |\theta_i| > \frac{\phi}{2} \sqrt{\frac{\log p}{n}} s_1^{\frac{1}{2-q}} \right\}, B_2'^{-} = \left\{ i, |\eta_i| > \frac{\phi}{2} \sqrt{\frac{\log p}{n}} s_2^{\frac{1}{2-q}} \right\}.$$

We need the bound of cardinality of $\left(B_i^{-} \right)^c - \left(B_i'^{-} \right)^c$. Following the lines of the proof of Lemma A.1, we have that

$$\left| \left(B_i^{-} \right)^c - \left(B_i'^{-} \right)^c \right| \leq C \left[s_i^{\left(1 - \frac{q}{2-q} \right)} \left(\frac{\log p}{n} \right)^{-q/2} + \xi_\Omega^2 \left(\frac{\log p}{n} \right)^{-1} s_i^{-\frac{2}{2-q}} \right], \qquad (28)$$

where we use the fact $\|\theta - \alpha\| \leq \left\| \hat{\Omega}_1 - \Omega_1 \right\| \|\Sigma_1 \alpha\| \leq C\xi_\Omega$ by the Assumption B. Now we bound $\left\| \alpha_{\left(B_1^{-} \right)^c} \right\|^2 = \sum_{i \in \left(B_1^{-} \right)^c} \alpha_i^2$ as follows,

$$\begin{aligned} \left\| \alpha_{\left(B_1^{-} \right)^c} \right\| &\leq \left\| \theta_{\left(B_1^{-} \right)^c} \right\| + C\xi_\Omega \leq \left\| \theta_{\left(B_1^{-} \right)^c - \left(B_1'^{-} \right)^c} \right\| + \left\| \theta_{\left(B_1'^{-} \right)^c} \right\| + C\xi_\Omega \\ &\leq \left\| \alpha_{\left(B_1^{-} \right)^c - \left(B_1'^{-} \right)^c} \right\| + \left\| \theta_{\left(B_1'^{-} \right)^c} \right\| + 2C\xi_\Omega \\ &\leq \left[C_1 \left(\frac{\log p}{n} \right)^{1-q/2} s_1^2 + C_2 \xi_\Omega^2 \right]^{1/2}, \end{aligned}$$

since we can bound the first two terms by Equation (28) and weak l_q ball assumption on θ as follows,

$$\begin{aligned} \left\| \alpha_{\left(B_1^{-} \right)^c - \left(B_1'^{-} \right)^c} \right\|^2 &\leq \phi^2 \frac{\log p}{n} s_1^{\frac{2}{2-q}} \left| \left(B_i^{-} \right)^c - \left(B_i'^{-} \right)^c \right| \leq C \left(\xi_\Omega^2 + s_1^2 \left(\frac{\log p}{n} \right)^{1-q/2} \right), \\ \left\| \theta_{\left(B_1'^{-} \right)^c} \right\|^2 &\leq \sum_{i \in \left(B_1'^{-} \right)^c} \theta_i^2 \leq \sum_k \left(\frac{\phi}{2} \right)^2 \frac{\log p}{n} s_1^{\frac{2}{2-q}} \wedge s_1^{2/q} k^{-2/q} \\ &\leq C \left(\frac{\log p}{n} \right)^{1-q/2} s_1^2. \end{aligned}$$

Similarly, we can obtain that $\left\| \beta_{\left(B_2^{-} \right)^c} \right\|^2 = \sum_{i \in \left(B_2^{-} \right)^c} \beta_i^2 \leq C_1 \left(\frac{\log p}{n} \right)^{1-q/2} s_2^2 + C_2 \xi_\Omega^2$. Therefore we finished the proof of Equation (19) and the lemma.

C. Proof of Lemma 7.1

In this section, we are going to show that the first pair of singular vectors of $\hat{A}^{ora}$ is close to the first pair of singular vectors of $A = \lambda \alpha \beta^T$. We introduce an intermediate step by involving another matrix

$$A^{ora} = \begin{pmatrix} \lambda \alpha_{H_1} \beta_{H_2}^T & 0 \\ 0 & 0 \end{pmatrix}.$$

It is easy to see $(\alpha/||\alpha||, \beta/||\beta||)$ is the first pair of singular vectors of A and $(\alpha^{ora}/||\alpha^{ora}||, \beta^{ora}/||\beta^{ora}||)$ is the first pair of singular vectors of A^{ora}, where

$$\alpha^{ora} = \begin{pmatrix} \alpha_{H_1} \\ 0 \end{pmatrix}, \quad \beta^{ora} = \begin{pmatrix} \beta_{H_2} \\ 0 \end{pmatrix}.$$

Let $(\hat{\alpha}^{ora}, \hat{\beta}^{ora})$ be the first pair of singular vectors of $\hat{A}^{ora}$. Then, we have

$$L(\hat{\alpha}^{ora}, \alpha) \le L(\hat{\alpha}^{ora}, \alpha^{ora}) + L(\alpha^{ora}, \alpha).$$

We have a similar inequality for β. We present a deterministic lemma before proving the results.

LEMMA C.1. *We have*

$$
\begin{aligned}
L(\hat{\alpha}^{ora}, \alpha) &\le \frac{\sqrt{2}||\hat{A}_{H_1 H_2} - A_{H_1 H_2}||}{\hat{l}_1} + \frac{8\sqrt{2}||\theta - \alpha||}{0.9W^{-1/2}} \\
&\quad + \frac{2\sqrt{2}}{0.9W^{-1/2}} \left(2^{\frac{q+1}{2}} + \sqrt{\frac{2}{2-q}} \right) \left(\delta_1^{2-q} s_1 \left(\frac{\log p_1}{n} \right)^{1-q/2} \right)^{1/2}, \\
L(\hat{\beta}^{ora}, \beta) &\le \frac{\sqrt{2}||\hat{A}_{H_1 H_2} - A_{H_1 H_2}||}{\hat{l}_1} + \frac{8\sqrt{2}||\eta - \beta||}{0.9W^{-1/2}} \\
&\quad + \frac{2\sqrt{2}}{0.9W^{-1/2}} \left(2^{\frac{q+1}{2}} + \sqrt{\frac{2}{2-q}} \right) \left(\delta_2^{2-q} s_2 \left(\frac{\log p_2}{n} \right)^{1-q/2} \right)^{1/2}.
\end{aligned}
$$

PROOF. Starting with the above triangle inequality, we need to bound $L(\hat{\alpha}^{ora}, \alpha^{ora})$ and $L(\alpha^{ora}, \alpha)$. For the first term, we use Wedin's sin-theta theorem (Theorem 4.4 in Stewart and Sun (1990)).

$$
\begin{aligned}
L(\hat{\alpha}^{ora}, \alpha^{ora}) &\le \sqrt{2} \left\| \frac{\hat{\alpha}^{ora} \hat{\alpha}^{ora,T}}{||\hat{\alpha}^{ora}||^2} - \frac{\alpha^{ora} \alpha^{ora,T}}{||\alpha^{ora}||^2} \right\| \\
&\le \frac{\sqrt{2} \left\| \hat{A}_{H_1 H_2} \frac{\beta_{H_2}}{||\beta_{H_2}||} - \lambda ||\beta_{H_2}|| \alpha_{H_1} \right\|}{\hat{l}_1} \\
&\le \frac{\sqrt{2}||\hat{A}_{H_1 H_2} - A_{H_1 H_2}||}{\hat{l}_1}, \tag{29}
\end{aligned}
$$

where we also applied the fact $l_2 = 0$ in Lemma A.6. For the second term, we have

$$
\begin{aligned}
L(\alpha^{ora}, \alpha) &\le \sqrt{2} \left\| \frac{\alpha^{ora}}{||\alpha^{ora}||} - \frac{\alpha}{||\alpha||} \right\| \\
&\le \frac{2\sqrt{2}||\alpha - \alpha^{ora}||}{||\alpha||} \\
&\le \frac{2\sqrt{2}||\alpha_{L_1}||}{0.9W^{-1/2}}, \tag{30}
\end{aligned}
$$

where the last inequality follows from Assumption B which leads to $||\alpha|| = (1 + o(1)) \, ||\theta|| \geq 0.9||\theta|| \geq 0.9W^{-1/2}$. Notice that

$$
\begin{aligned}
||\alpha_{L_1}|| &\leq ||\theta_{L_1}|| + ||\theta - \alpha|| \\
&\leq ||\theta_{L_1 - L_1'}|| + ||\theta_{L_1'}|| + ||\theta - \alpha|| \\
&\leq ||\alpha_{L_1 - L_1'}|| + ||\theta_{L_1'}|| + 2||\theta - \alpha|| \\
&\leq \left(2^{\frac{1+q}{2}} + \sqrt{\frac{2}{2-q}} \right) \left(\delta_1^{2-q} s_1 \left(\frac{\log p_1}{n} \right)^{1-q/2} \right)^{1/2} + 4||\theta - \alpha||,
\end{aligned}
$$

because $||\alpha_{L_1 - L_1'}||^2 \leq \delta_1^2 \frac{\log p_1}{n} |L_1 - L_1'| \leq 2^{1+q} \delta_1^{2-q} s_1 \left(\frac{\log p_1}{n} \right)^{1-q/2} + 4||\theta - \alpha||^2$, and

$$
\begin{aligned}
||\theta_{L_1'}||^2 &= \sum_{k \in L_1'} \theta_k^2 \\
&\leq \sum_{k=1}^{p_1} \left((\delta_1/2)^2 \frac{\log p_1}{n} \wedge s_1^{2/q} k^{-2/q} \right) \\
&\leq \int_0^\infty \left((\delta_1/2)^2 \frac{\log p_1}{n} \right) \wedge s_1^{2/q} x^{-2/q} dx \\
&\leq (\delta_1/2)^{2-q} s_1 \left(\frac{\log p_1}{n} \right)^{1-q/2} + \frac{q}{2-q} (\delta_1/2)^{2-q} s_1 \left(\frac{\log p_1}{n} \right)^{1-q/2} \\
&\leq \frac{2}{2-q} \delta_1^{2-q} s_1 \left(\frac{\log p_1}{n} \right)^{1-q/2}.
\end{aligned}
$$

Combining Equations (29) and (30), together with the bounds above, we finish our proof for $L(\hat{\alpha}^{ora}, \alpha)$. Similar analysis works for $L(\hat{\beta}^{ora}, \beta)$ and we omit the details.

PROOF (PROOF OF LEMMA 7.1). The proof directly follows from the deterministic bound in Lemma C.1 and the probabilistic bound in Lemma A.6.

D. Proofs of Lemma 7.2 and Lemma 7.3

In this section, we are going to give the convergence bound for the Algorithms 1 and 2 applied on the oracle matrix $\hat{A}^{ora}$. According to Lemma B.1, we obtain that the initializer applied to the oracle matrix $\hat{A}^{ora}$ by Algorithm 2 is identical to that applied to data matrix $\hat{A}$, i.e. $(\alpha^{(0)}, \beta^{(0)}) = (\alpha^{(0),ora}, \beta^{(0),ora})$ since $B_i \subset H_i$. As a consequence, the properties in Lemma B.1 also hold for $(\alpha^{(0),ora}, \beta^{(0),ora})$.

The following three lemmas are helpful for us to understand the convergence of oracle sequence $(\alpha^{(k),ora}, \beta^{(k),ora})$ obtained from Algorithm 1. By definition, the initializer satis-

fies $\alpha_{L_1}^{(0),ora} = 0$ and $\beta_{L_2}^{(0),ora} = 0$. Then, Algorithm 1 only involves the sub-matrix $\hat{A}_{H_1 H_2}$. We first state the basic one-step analysis for power SVD method in the following lemma.

LEMMA D.1. *Let A be a matrix with first and second eigenvalues (l_1, l_2) satisfying $|l_1| > |l_2|$. Let (u, v) be the first pair of singular vectors and (α, β) be any vectors. Define $\bar{\alpha} = A\beta$ and $\bar{\beta} = A^T \alpha$. We have*

$$\frac{1}{2}L(\bar{\alpha}, u)^2 \leq \frac{\frac{1}{2}L(\beta, v)^2}{1 - \frac{1}{2}L(\beta, v)^2}\left|\frac{l_2}{l_1}\right|^2,$$

$$\frac{1}{2}L(\bar{\beta}, v)^2 \leq \frac{\frac{1}{2}L(\alpha, v)^2}{1 - \frac{1}{2}L(\alpha, v)^2}\left|\frac{l_2}{l_1}\right|^2.$$

PROOF. We omit the proof because it is almost identical to the proof of Theorem 8.2.2 in Golub and Van Loan (1996).

Applying the above lemma in our case on the matrix $\hat{A}_{H_1 H_2}$ and with unit-vector initializers $(\alpha_{H_1}^{(0)}, \beta_{H_2}^{(0)})$. For simplicity of notations, we drop the subscript and write $\hat{A}$ and $(\alpha^{(0)}, \beta^{(0)})$ in this section. At the k-th step, we write

$$\bar{\alpha}^{(k)} = \hat{A}\beta^{(k-1)}, \quad \bar{\beta}^{(k)} = \hat{A}^T\alpha^{(k-1)},$$

$$\alpha^{(k)} = \frac{T(\bar{\alpha}^{(k)}, \gamma_1)}{||T(\bar{\alpha}^{(k)}, \gamma_1)||}, \quad \beta^{(k)} = \frac{T(\bar{\beta}^{(k)}, \gamma_2)}{||T(\bar{\beta}^{(k)}, \gamma_2)||}.$$

Now we give a one-step analysis for Algorithm 1.

LEMMA D.2. *Let $(\hat{\alpha}, \hat{\beta})$ be the first pair of singular vectors of $\hat{A}$ and $(\hat{l}_1, \hat{l}_2)$ be the first and second singular values. We assume*

$$\frac{1}{2}L(\alpha^{(0)}, \hat{\alpha})^2 \leq \frac{1}{2}, \quad \frac{1}{2}L(\beta^{(0)}, \hat{\beta})^2 \leq \frac{1}{2},$$

and

$$\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{8\left(\gamma_1^2|H_1| \vee \gamma_2^2|H_2|\right)}{|\hat{l}_1|^2} \leq \frac{1}{4}.$$

Then, we have

$$\frac{1}{4}L(\alpha^{(k)}, \hat{\alpha})^2 \leq L(\beta^{(k-1)}, \hat{\beta})^2\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{8\gamma_1^2|H_1|}{|\hat{l}_1|^2},$$

$$\frac{1}{4}L(\beta^{(k)}, \hat{\beta})^2 \leq L(\alpha^{(k-1)}, \hat{\alpha})^2\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{8\gamma_2^2|H_2|}{|\hat{l}_1|^2}.$$

PROOF. By triangle inequality, we have

$$L(\alpha^{(k)}, \hat{\alpha}) \leq L(\alpha^{(k)}, \bar{\alpha}^{(k)}) + L(\bar{\alpha}^{(k)}, \hat{\alpha}).$$

The second term on the right hand side above is bounded in Lemma D.1. The first term is bounded as

$$
\begin{aligned}
L(\alpha^{(k)}, \bar{\alpha}^{(k)}) &\leq \frac{2\sqrt{2}||T(\bar{\alpha}^{(k)}, \gamma_1) - \bar{\alpha}^{(k)}||}{||\bar{\alpha}^{(k)}||} \\
&\leq \frac{2\sqrt{2}\gamma_1\sqrt{|H_1|}}{||\hat{A}\beta^{(k-1)}||},
\end{aligned}
$$

where $||\hat{A}\beta^{(k-1)}||^2 \geq |\hat{l}_1|^2|\hat{\beta}^T\beta^{(k-1)}|^2 = |\hat{l}_1|^2\left(1 - \frac{1}{2}L(\beta^{(k-1)}, \hat{\beta})^2\right)$. Therefore, we have

$$
\frac{1}{4}L(\alpha^{(k)}, \hat{\alpha})^2 \leq \frac{\frac{1}{2}L(\beta^{(k-1)}, \hat{\beta})^2}{1 - \frac{1}{2}L(\beta^{(k-1)}, \hat{\beta})^2}\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{4\gamma_1^2|H_1|/|\hat{l}_1|^2}{1 - \frac{1}{2}L(\beta^{(k-1)}, \hat{\beta})^2}.
$$

In the same way, we have

$$
\frac{1}{4}L(\beta^{(k)}, \hat{\beta})^2 \leq \frac{\frac{1}{2}L(\alpha^{(k-1)}, \hat{\alpha})^2}{1 - \frac{1}{2}L(\alpha^{(k-1)}, \hat{\alpha})^2}\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2 + \frac{4\gamma_2^2|H_2|/|\hat{l}_1|^2}{1 - \frac{1}{2}L(\alpha^{(k-1)}, \hat{\alpha})^2}.
$$

Suppose $L(\alpha^{(k-1)}, \hat{\alpha})^2 \vee L(\beta^{(k-1)}, \hat{\beta})^2 \leq 1$, then it is easy to see that $L(\alpha^{(k)}, \hat{\alpha})^2 \vee L(\beta^{(k)}, \hat{\beta})^2 \leq 1$ under the assumption. Using mathematical induction, $L(\alpha^{(k-1)}, \hat{\alpha})^2 \vee L(\beta^{(k-1)}, \hat{\beta})^2 \leq 1$ is true for each k. Therefore, we deduce the desired result.

The above Lemma D.2 implies the following convergence rate of oracle sequence.

LEMMA D.3. *Suppose the assumptions of Lemma D.2 hold, and we further assume* $|\hat{l}_2|/|\hat{l}_1| < 32^{-1/4}$, *and then we have for all* $k \geq 2$,

$$
L(\alpha^{(k)}, \hat{\alpha})^2 \leq \max\left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2\frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2} + \frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2}, \left(32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4\right)^{[k/2]}\right),
$$

$$
L(\beta^{(k)}, \hat{\beta})^2 \leq \max\left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2\frac{64\gamma_1^2|H_1|}{|\hat{l}_1|^2} + \frac{64\gamma_2^2|H_2|}{|\hat{l}_1|^2}, \left(32\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^4\right)^{[k/2]}\right).
$$

PROOF. We only prove the bound for $L(\alpha^{(k)}, \hat{\alpha})$. Using the previous lemma, we derive a formula of a two-step analysis

$$
L(\alpha^{(k)}, \hat{\alpha})^2 \leq L(\alpha^{(k-2)}, \hat{\alpha})^2\rho + \omega_1,
$$

where

$$
\rho = \left(4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2\right)^2, \quad \omega_1 = 4\left|\frac{\hat{l}_2}{\hat{l}_1}\right|^2\frac{32\gamma_2^2|H_2|}{|\hat{l}_1|^2} + \frac{32\gamma_1^2|H_1|}{|\hat{l}_1|^2}.
$$

Therefore, for each k, we have

$$
L(\alpha^{(k)}, \hat{\alpha})^2 \leq \max\left(2\omega_1, L(\alpha^{(k-2)}, \hat{\alpha})^2(2\rho)\right).
$$

We are going to prove for each k,

$$L(\alpha^{(2k)}, \hat{\alpha})^2 \le \max\left(2\omega_1, L(\alpha^{(0)}, \hat{\alpha})^2(2\rho)^k\right).$$

It is obvious that this is true for $k = 0$. Suppose this is true for $k - 1$, then we have

$$
\begin{aligned}
L(\alpha^{(2k)}, \hat{\alpha})^2 &\le \max\left(2\omega_1, L(\alpha^{(2(k-1))}, \hat{\alpha})^2(2\rho)\right) \\
&\le \max\left(2\omega_1, \max\left(2\omega_1, L(\alpha^{(0)}, \hat{\alpha})^2(2\rho)^{k-1}\right)(2\rho)\right) \\
&\le \max\left(2\omega_1, 2\omega_1(2\rho), L(\alpha^{(0)}, \hat{\alpha})^2(2\rho)^k\right) \\
&= \max\left(2\omega_1, L(\alpha^{(0)}, \hat{\alpha})^2(2\rho)^k\right),
\end{aligned}
$$

where the last inequality follows from the assumption $|\hat{l}_2/\hat{l}_1| \le 32^{-1/4}$. By mathematical induction, the inequality is true for each k. Similarly, we can show that for each k,

$$L(\alpha^{(2k+1)}, \hat{\alpha})^2 \le \max\left(2\omega_1, L(\alpha^{(1)}, \hat{\alpha})^2(2\rho)^k\right).$$

Therefore,

$$L(\alpha^{(k)}, \hat{\alpha})^2 \le \max\left(2\omega_1, (2\rho)^{[k/2]}\right),$$

and the proof is complete by a similar argument for $\beta^{(k)}$.

PROOF (PROOF OF LEMMA 7.2). It is sufficient to check the conditions of Lemma D.2 and Lemma D.3. The first condition of Lemma D.2

$$\frac{1}{2}L^2(\alpha^{(0)}, \hat{\alpha})^2 \le \frac{1}{2}, \quad \frac{1}{2}L(\beta^{(0)}, \hat{\beta})^2 \le \frac{1}{2},$$

is directly by the fact that the initializer is consistent, which is guaranteed by Lemma B.1. The second condition of Lemma D.2 is deduced from Lemma A.6 (bounds of $|\hat{l}_1|$ and $|\hat{l}_2|$) and Lemma A.1 (bounds of $|H_1|$ and $|H_2|$). Finally, the condition of Lemma D.3 follows from Lemma A.6. The conclusions of Lemma 7.2 are the conclusions of Lemma D.3 and Lemma D.2 respectively.

PROOF (PROOF OF LEMMA 7.3). The bounds on $|H_1|$ and $|H_2|$ have been established in Lemma A.1. The bounds for $|\hat{l}_2|^2$ and $|\hat{l}_1|^{-2}$ are from Lemma A.6.

E. Proofs of Lemma 7.4 and Lemma 7.5

We first present a deterministic bound and then prove the results by probabilistic arguments.

LEMMA E.1. *For any unit vectors $a \in \mathbb{R}^{|H_1|}$ and $b \in \mathbb{R}^{|H_2|}$, we have*

$$
\begin{aligned}
||\hat{A}_{L_1 H_2} b||_\infty &\leq \frac{1}{n}\sum_{i=1}^n Z_i^2 ||\beta|| \delta_1 \sqrt{\frac{\log p_1}{n}} + \delta_1 \sqrt{\frac{\log p_1}{n}} \left\|\frac{1}{n}\sum_{i=1}^n Z_i Y_{i,H_2}'^T \right\| + ||\beta|| \left\|\frac{1}{n}\sum_{i=1}^n Z_i X_{i,L_1}' \right\|_\infty \\
&\quad + \frac{||Y_{H_2}'^T Y_{H_2}'||^{1/2}}{n} \max_{k \in L_1} \sqrt{\mathrm{Var}(X_{1,k}')} \left| \frac{\sum_{i=1}^n X_{i,k}' Y_{i,H_2}'^T b}{\sqrt{\mathrm{Var}(X_{1,k}') b^T Y_{H_2}'^T Y_{H_2}' b}} \right|,
\end{aligned}
$$

$$
\begin{aligned}
||\hat{A}_{H_1 L_2}^T a||_\infty &\leq \frac{1}{n}\sum_{i=1}^n Z_i^2 ||\alpha|| \delta_2 \sqrt{\frac{\log p_2}{n}} + \delta_2 \sqrt{\frac{\log p_2}{n}} \left\|\frac{1}{n}\sum_{i=1}^n Z_i X_{i,H_1}'^T \right\| + ||\alpha|| \left\|\frac{1}{n}\sum_{i=1}^n Z_i Y_{i,L_2}' \right\|_\infty \\
&\quad + \frac{||X_{H_1}'^T X_{H_1}'||^{1/2}}{n} \max_{k \in L_2} \sqrt{\mathrm{Var}(Y_{1,k}')} \left| \frac{\sum_{i=1}^n Y_{i,k}' X_{i,H_1}^T a}{\sqrt{\mathrm{Var}(Y_{1,k}') a^T Y_{H_1}'^T Y_{H_1}' a}} \right|.
\end{aligned}
$$

PROOF. Using the latent representation in Lemma A.2, we have

$$
\tilde{X}_{i,L_1} \tilde{Y}_{i,H_2}^T = \lambda Z_i^2 \alpha_{L_1} \beta_{H_2}^T + \sqrt{\lambda} Z_i \alpha_{L_1} Y_{i,H_2}'^T + \sqrt{\lambda} Z_i X_{i,L_1}' \beta_{H_2}^T + X_{i,L_1}' Y_{i,H_2}'^T.
$$

Therefore,

$$
\begin{aligned}
||\hat{A}_{L_1 H_2} b||_\infty &\leq \lambda \frac{1}{n}\sum_{i=1}^n Z_i^2 ||\alpha_{L_1}||_\infty |\beta_{H_2}^T b| + \sqrt{\lambda} ||\alpha_{L_1}||_\infty \left|\frac{1}{n}\sum_{i=1}^n Z_i Y_{i,H_2}'^T b\right| \\
&\quad + \sqrt{\lambda} \left\|\frac{1}{n}\sum_{i=1}^n Z_i X_{i,L_1}' \right\|_\infty |\beta_{H_2}^T b| + \left\|\frac{1}{n}\sum_{i=1}^n X_{i,L_1}' Y_{i,H_2}'^T b\right\|_\infty,
\end{aligned}
$$

where the first term is bounded by

$$
\frac{1}{n}\sum_{i=1}^n Z_i^2 ||\beta|| \delta_1 \sqrt{\frac{\log p_1}{n}},
$$

the second term is bounded by

$$
\delta_1 \sqrt{\frac{\log p_1}{n}} \left\|\frac{1}{n}\sum_{i=1}^n Z_i Y_{i,H_2}'^T \right\|,
$$

the third term is bounded by

$$
||\beta|| \left\|\frac{1}{n}\sum_{i=1}^n Z_i X_{i,L_1}' \right\|_\infty,
$$

and the last term is bounded by

$$
\begin{aligned}
\max_{k \in L_1} \left|\frac{1}{n}\sum_{i=1}^n X_{i,k}' Y_{i,H_2}^T b\right| &\leq \max_{k \in L_1} \sqrt{\mathrm{Var}(X_{1,k}')} \left|\frac{1}{n}\sum_{i=1}^n \frac{X_{i,k}'}{\sqrt{\mathrm{Var}(X_{1,k}')}} Y_{i,H_2}'^T b\right| \\
&\leq \frac{||Y_{H_2}'^T Y_{H_2}'||^{1/2}}{n} \max_{k \in L_1} \sqrt{\mathrm{Var}(X_{1,k}')} \left| \frac{\sum_{i=1}^n X_{i,k}' Y_{i,H_2}^T b}{\sqrt{\mathrm{Var}(X_{1,k}') b^T Y_{H_2}'^T Y_{H_2}' b}} \right|,
\end{aligned}
$$

where Y_{H_2}' is an $n \times |H_2|$ matrix with the i-th row $Y_{i,H_2}'^T$. Summing up the bounds, the proof is complete.

PROOF (PROOF OF LEMMA 7.4). This is a corollary of Result 1 of Lemma B.1. By $B_i \subset H_i$ for $i = 1, 2$, we have $B_i^{ora} = B_i$ for $i = 1, 2$. Thus, $(\alpha^{(0),ora}, \beta^{(0),ora}) = (\alpha^{(0)}, \beta^{(0)})$.

PROOF (PROOF OF LEMMA 7.5). We upper bound each term in the conclusion of Lemma E.1. According to the concentration inequalities we have established,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^2 \le 2,$$

with probability at least $1 - 2e^{-3n/16}$ by Lemma A.3.

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Y_{i,H_2}^{\prime T}\right\| \le 2.06||\hat{\Omega}_2||^{1/2},$$

with probability at least $1 - \left(e^{-3n/64} + e^{-(\sqrt{n}-1)^2/2}\right)$ by Lemma A.5.

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i X_{i,L_1}^{\prime}\right\|_{\infty} \le 2.07||\hat{\Omega}_1||^{1/2}\sqrt{\frac{\log p}{n}},$$

with probability at least $1 - \left(e^{-3n/64+\log p_1} + p^{-2}\right)$ by Lemma A.5.

$$\frac{||Y_{H_2}^{\prime T}Y_{H_2}^{\prime}||^{1/2}}{n} \le 3.02||\hat{\Omega}_2||^{1/2}n^{-1/2},$$

with probability at least $1 - 2e^{-n/2}$ by Lemma A.4. Using union bound and by Lemma E.1, we have

$$
\begin{aligned}
||\hat{A}_{L_1 H_2} b^{(k)}||_{\infty} \le\ & \left(2\delta_1||\beta|| + 2.06\delta_1||\hat{\Omega}_2||^{1/2} + 2.07||\beta||||\hat{\Omega}_1||^{1/2}\right)\sqrt{\frac{\log p}{n}} \\
& + 3.02||\hat{\Omega}_2||^{1/2}n^{-1/2}\max_{l \in L_1}\sqrt{\mathrm{Var}(X_{1,l}^{\prime})}\left|\frac{\sum_{i=1}^{n} X_{i,l}^{\prime}Y_{i,H_2}^{\prime T}b^{(k)}}{\sqrt{\mathrm{Var}(X_{1,l}^{\prime})b^{(k),T}Y_{H_2}^{\prime T}Y_{H_2}^{\prime}b^{(k)}}}\right|,
\end{aligned}
$$

with probability at least $1 - O(p^{-2})$ for all k. Notice by Lemma A.2

$$\max_{l \in L_1}\sqrt{\mathrm{Var}(X_{1,l}^{\prime})}\left|\frac{\sum_{i=1}^{n} X_{i,l}^{\prime}Y_{i,H_2}^{\prime T}b^{(k)}}{\sqrt{\mathrm{Var}(X_{1,l}^{\prime})b^{(k),T}Y_{H_2}^{\prime T}Y_{H_2}^{\prime}b^{(k)}}}\right| \le 1.01||\hat{\Omega}_2||^{1/2}\max_{l \in L_1}\left|\frac{\sum_{i=1}^{n} X_{i,l}^{\prime}Y_{i,H_2}^{\prime T}b^{(k)}}{\sqrt{\mathrm{Var}(X_{1,l}^{\prime})b^{(k),T}Y_{H_2}^{\prime T}Y_{H_2}^{\prime}b^{(k)}}}\right|.$$

Since $b^{(k)}$ only depends on $\hat{A}_{H_1 H_2}$, $Y_{H_2}^{\prime}$ and $b^{(k)}$ are jointly independent of $X_{L_1}^{\prime}$. Therefore, conditioning on Y_{H_2} and $b^{(k)}$,

$$\frac{\sum_{i=1}^{n} X_{i,l}^{\prime}Y_{i,H_2}^{\prime T}b^{(k)}}{\sqrt{\mathrm{Var}(X_{1,l}^{\prime})b^{(k),T}Y_{H_2}^{\prime T}Y_{H_2}^{\prime}b^{(k)}}}$$

is a standard Gaussian. Therefore, by union bound,

$$\max_{l \in L_1}\left|\frac{\sum_{i=1}^{n} X_{i,l}^{\prime}Y_{i,H_2}^{\prime T}b^{(k)}}{\sqrt{\mathrm{Var}(X_{1,l}^{\prime})b^{(k),T}Y_{H_2}^{\prime T}Y_{H_2}^{\prime}b^{(k)}}}\right| \le \sqrt{6\log p},$$

with probability at least $1 - Kp^{-2}$, for all $k = 1, 2..., K$. Finally we have

$$||\hat{A}_{L_1 H_2} b^{(k)}||_\infty \leq \left(2\delta_1||\beta|| + 2.06\delta_1||\hat{\Omega}_2||^{1/2} + 2.07||\beta||||\hat{\Omega}_1||^{1/2} + 7.5||\hat{\Omega}_2||\right)\sqrt{\frac{\log p}{n}},$$

for all $k = 1, 2, ..., K$, with probability at least $1 - O(p^{-2})$. The same analysis also applies to $||\hat{A}_{H_1 L_2}^T a^{(k)}||_\infty$. The result is proved by $||\alpha|| \leq 1.01||\hat{\Omega}_1||^{1/2}$ and $||\beta|| \leq 1.01||\hat{\Omega}_2||^{1/2}$ and the choice of γ_1 and γ_2 in Section 7.2.

F.   Proofs of Technical Lemmas

PROOF (PROOF OF LEMMA A.1). By definition,

$$
\begin{aligned}
|H_1'| &\leq \left|\left\{k : |\theta_k| \geq \frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right\}\right| \\
&= \left|\left\{k : |\theta_{(k)}| \geq \frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right\}\right| \\
&\leq \left|\left\{k : s_1 k^{-1} \geq \left(\frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right)^q\right\}\right| \\
&\leq (\delta_1/2)^{-q} s_1 \left(\frac{\log p_1}{n}\right)^{-q/2}.
\end{aligned}
$$

Notice

$$
\begin{aligned}
H_1 &\subset \left\{k : |\theta_k| + |\alpha_k - \theta_k| \geq \delta_1\sqrt{\frac{\log p_1}{n}}\right\} \\
&\subset H_1' \bigcup \left\{k : |\alpha_k - \theta_k| \geq \frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right\}.
\end{aligned}
$$

Since

$$\left|\left\{k : |\alpha_k - \theta_k| \geq \frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right\}\right|\left(\frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right)^2 \leq ||\alpha - \theta||^2,$$

we have

$$
\begin{aligned}
|H_1| &\leq |H_1'| + \left|\left\{k : |\alpha_k - \theta_k| \geq \frac{1}{2}\delta_1\sqrt{\frac{\log p_1}{n}}\right\}\right| \\
&\leq (\delta_1/2)^{-q} s_1 \left(\frac{\log p_1}{n}\right)^{-q/2} + (\delta_1/2)^{-2}\left(\frac{\log p_1}{n}\right)^{-1}||\theta - \alpha||^2.
\end{aligned}
$$

We also have

$$|L_1 - L_1'| = |H_1' - H_1| \leq |H_1| + |H_1'| \leq 2(\delta_1/2)^{-q} s_1 \left(\frac{\log p_1}{n}\right)^{-q/2} + (\delta_1/2)^{-2}\left(\frac{\log p_1}{n}\right)^{-1}||\theta - \alpha||^2.$$

Similar arguments apply to $|H_2|, |H_2'|, |H_2 - H_2'|, |L_2 - L_2'|$.

PROOF (PROOF OF LEMMA A.2). It is not hard to find the formula

$$\text{Cov}(X') = \hat{\Omega}_1 \Sigma_1 \hat{\Omega}_1 - \lambda \alpha \alpha^T, \quad \text{Cov}(Y') = \hat{\Omega}_2 \Sigma_2 \hat{\Omega}_2 - \lambda \beta \beta^T.$$

Plugging $\alpha = \hat{\Omega}_1 \Sigma_1 \theta$, we have

$$\text{Cov}(X') = \hat{\Omega}_1 \Sigma_1^{1/2} \left(I - \lambda \Sigma_1^{1/2} \theta \theta^T \Sigma_1^{1/2} \right) \Sigma_1^{1/2} \hat{\Omega}_1.$$

Since $||\Sigma^{1/2}\theta|| = 1$ and $\lambda \geq 1$, we have $\text{Cov}(X') \geq 0$. We proceed to prove the spectral bound as follows.

$$||\text{Cov}(X')|| \leq ||\hat{\Omega}_1 \Sigma_1 \hat{\Omega}_1|| \leq \left(1 + o(1) \right) ||\hat{\Omega}_1||,$$

where the last inequality follows from Equation (11) in Assumption B. The same results also hold for $\text{Cov}(Y')$.

PROOF (PROOF OF LEMMA A.4). Let us denote the covariance matrix of each row of the matrix X'_{H_1} by R. Then, we have $||X'^T_{H_1} X'_{H_1}|| \leq ||R|| ||U^T_{H_1} U_{H_1}||$, where U_{H_1} is an $n \times |H_1|$ Gaussian random matrix. We bound $||R||$ according to Lemma A.2 by

$$||R|| \leq ||\text{Cov}(X')|| \leq 1.01 ||\hat{\Omega}_1||.$$

For $||U^T_{H_1} U_{H_1}||$, we have the bound

$$\mathbb{P}\left(\frac{1}{n} ||U^T_{H_1} U_{H_1}|| - 1 > 2 \left(\sqrt{\frac{|H_1|}{n}} + t \right) + \left(\sqrt{\frac{|H_1|}{n}} + t \right)^2 \right)$$
$$\leq \mathbb{P}\left(||\frac{1}{n} U^T_{H_1} U_{H_1} - I|| > 2 \left(\sqrt{\frac{|H_1|}{n}} + t \right) + \left(\sqrt{\frac{|H_1|}{n}} + t \right)^2 \right)$$
$$\leq 2e^{-nt^2/2},$$

where the last inequality is from Proposition D.1 in Ma (2013). In the similar way, we obtain the bound for $||Y'^T_{H_2} Y'_{H_2}||$. Now we bound $||X'^T_{H_1} Y'_{H_2}||$. Denote the covariance of each row of the matrix Y'_{H_2} by S. Then we have $||X'^T_{H_1} Y'_{H_2}|| \leq ||R||^{1/2} ||S||^{1/2} ||U^T_{H_1} V_{H_2}||$, where V_{H_2} is an $n \times |H_2|$ Gaussian random matrix. We have $||R||^{1/2} ||S||^{1/2} \leq 1.01 ||\hat{\Omega}_1||^{1/2} ||\hat{\Omega}_2||^{1/2}$ by Lemma A.2, and

$$\mathbb{P}\left(||U^T_{H_1} V_{H_2}|| > 1.01 \left(\sqrt{|H_1|n} + \sqrt{|H_2|n} + t\sqrt{n} \right) \right) \leq \left(|H_1| \wedge |H_2| \right) e^{-3n/64} + e^{-t^2/2},$$

from Proposition D.2 in Ma (2013). Thus, the proof is complete.

PROOF (PROOF OF LEMMA A.5). Define Z be the vector $(Z_1, ..., Z_n)^T$. We keep the notations in the proof of the above lemma. Then, we have

$$\left\| \sum_{i=1}^n Z_i X'_{i,H_1} \right\| = ||Z^T X'_{H_1}|| \leq 1.01 ||\hat{\Omega}_1||^{1/2} ||Z^T U_{H_1}||,$$

where $||Z^T U_{H_1}||$ is upper bounded by

$$\mathbb{P}\left(||Z^T U_{H_1}|| > 1.01\Big((t+1)\sqrt{n} + \sqrt{|H_1|n}\Big)\right) \le e^{-3n/64} + e^{-t^2/2},$$

by Proposition D.2 in Ma (2013). The similar analysis also applies to $\left\|\sum_{i=1}^n Z_i Y'_{i,H_2}\right\|$. For the third inequality, we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Z_i X'_{i,L_1}\right\|_\infty > 1.03||\hat\Omega_1||^{1/2}(t+2)\sqrt{n}\right)$$

$$\le \sum_{k \in L_1} \mathbb{P}\left(\left|\sum_{i=1}^n Z_i X'_{ik}\right| > 1.03||\hat\Omega_1||^{1/2}(t+2)\sqrt{n}\right)$$

$$\le \sum_{k \in L_1} \mathbb{P}\left(\left|\sum_{i=1}^n Z_i \frac{X'_{ik}}{\sqrt{\mathrm{Var}(X'_{ik})}}\right| > 1.01(t+2)\sqrt{n}\right)$$

$$\le |L_1|\Big(e^{-3n/64} + e^{-t^2/2}\Big),$$

where we have used Proposition D.2 in Ma (2013) again. Similarly, we obtain the last inequality. The proof is complete.

PROOF (PROOF OF LEMMA A.6). Using the latent representation in Lemma A.2, we have

$$\hat A_{H_1 H_2} = A_{H_1 H_2} + A_{H_1 H_2}\left(\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1\right) + \sqrt{\lambda}\alpha_{H_1}\left(\frac{1}{n}\sum_{i=1}^n Z_i Y'_{i,H_2}{}^T\right)$$

$$+ \sqrt{\lambda}\left(\frac{1}{n}\sum_{i=1}^n Z_i X'_{i,H_1}{}^T\right)\beta_{H_2}^T + \frac{1}{n}\sum_{i=1}^n X'_{i,H_1} Y'_{i,H_2}{}^T.$$

Therefore

$$||\hat A_{H_1 H_2} - A_{H_1 H_2}|| \le ||A_{H_1 H_2}||\left|\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1\right| + ||\alpha||\left\|\frac{1}{n}\sum_{i=1}^n Z_i Y'_{i,H_2}{}^T\right\|$$

$$+ ||\beta||\left\|\frac{1}{n}\sum_{i=1}^n Z_i X'_{i,H_1}{}^T\right\| + \left\|\frac{1}{n}\sum_{i=1}^n X'_{i,H_1} Y'_{i,H_2}{}^T\right\|. \quad (31)$$

Now we control the upper bounds of four terms above. By picking $t = s^{1/2}\left(\frac{\log p}{n}\right)^{1/2-q/4}$, Lemma A.3 implies

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1\right| \le s^{1/2}\left(\frac{\log p}{n}\right)^{1/2-q/4},$$

with probability at least $1 - O(p^{-2})$. Moreover, Lemma A.5 implies that

$$\left\|\frac{1}{n}\sum_{i=1}^n Z_i Y'_{i,H_2}{}^T\right\| \le C||\hat\Omega_2||^{1/2}\left(s^{1/2}\Big(\frac{\log p}{n}\Big)^{1/2-q/4} + \frac{\sqrt{|H_2|}}{\sqrt{n}}\right)$$

$$\le C||\hat\Omega_2||^{1/2}\left(s^{1/2}\Big(\frac{\log p}{n}\Big)^{1/2-q/4} + ||\beta - \eta||\right)$$

with probability at least $1 - O(p^{-2})$ and

$$
\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i {X'_{i,H_1}}^T \right\| &\leq C \|\hat{\Omega}_1\|^{1/2} \left(s^{1/2} \Big(\frac{\log p}{n} \Big)^{1/2 - q/4} + \frac{\sqrt{|H_1|}}{\sqrt{n}} \right) \\
&\leq C \|\hat{\Omega}_1\|^{1/2} \left(s^{1/2} \Big(\frac{\log p}{n} \Big)^{1/2 - q/4} + \|\alpha - \theta\| \right)
\end{aligned}
$$

with probability at least $1 - O(p^{-2})$, where we also used Lemma A.1 to control $|H_1|$ and $|H_2|$. Similarly, Lemma A.1 and Lemma A.4 imply

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} X'_{i,H_1} {Y'_{i,H_2}}^T \right\| \leq C \|\hat{\Omega}_1\|^{1/2} \|\hat{\Omega}_2\|^{1/2} \left(s^{1/2} \Big(\frac{\log p}{n} \Big)^{1/2 - q/4} + \|\alpha - \theta\| + \|\beta - \eta\| \right),
$$

with probability at least $1 - O(p^{-2})$. Besides, Assumption B guarantees that $\|A_{H_1 H_2}\|$, $\|\hat{\Omega}_1\|^{1/2}$, $\|\hat{\Omega}_2\|^{1/2}$, $\|\alpha\|$ and $\|\beta\|$ are bounded above by some constant. Equation (31), together with above bounds, completes our proof for $\|\hat{A}_{H_1 H_2} - A_{H_1 H_2}\|$. The bound for $\max_{i=1,2} |\hat{l}_i - l_i|$ directly follows from Wely's theorem,

$$
\max_{i=1,2} |\hat{l}_i - l_i| \leq \|\hat{A}_{H_1 H_2} - A_{H_1 H_2}\|.
$$

For the last result, it's clear that $A_{H_1 H_2} = \lambda \alpha_{H_1} \beta_{H_2}$ is of rank one with $l_2 = 0$ and $l_1 = \lambda \|\alpha_{H_1}\| \|\beta_{H_2}\| \geq \lambda (\|\alpha\| - \|\alpha_{L_1}\|)(\|\beta\| - \|\beta_{L_2}\|)$. Assumption B implies that $\|\alpha\| = (1 + o(1)) \|\theta\| \geq 0.9 W^{-1/2}$, $\|\beta\| \geq 0.9 W^{-1/2}$ and $\lambda \geq C_\lambda$. To finish our proof that l_1 is bounded below away from zero $l_1 \geq C^{-1}$, we only need to show that $\|\alpha_{L_1}\| = o(1)$ and $\|\beta_{L_2}\| = o(1)$. This can be seen from our previous results Equation (19) $\max \left\{ \left\| \alpha_{(B_1^-)^c} \right\|, \left\| \beta_{(B_2^-)^c} \right\| \right\} = o(1)$ and Equation (17) $L_i \subset (B_i^-)^c$.

G. Proof of Theorem 4.2

The main tool for our proof is the Fano's Lemma, which is based on multiple hypotheses testing argument. To introduce Fano's Lemma, we first need to introduce a few notations. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$ with density p and q with respect to a common dominating measure μ, write the Kullback-Leibler divergence as $K(\mathbb{P}, \mathbb{Q}) = \int p \log \frac{p}{q} d\mu$. The following lemma, which can be viewed as a version of Fano's Lemma, gives a lower bound for the minimax risk over the parameter set $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{m_*}\}$ with the loss function $d(\cdot, \cdot)$. See Tsybakov (2009), Section 2.6 for more detailed discussions.

LEMMA G.1 (FANO). *Let $\Omega = \{\omega_k : k = 0, \ldots, m_*\}$ be a parameter set, where d is a distance over Ω. Let $\{\mathbb{P}_\omega : \omega \in \Omega\}$ be a collection of probability distributions satisfying*

$$
\frac{1}{m_*} \sum_{1 \leq k \leq m_*} K(\mathbb{P}_{\omega_k}, \mathbb{P}_{\omega_0}) \leq c \log m_* \tag{32}
$$

with $0 < c < 1/8$. Let $\hat{\omega}$ be any estimator based on an observation from a distribution in $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Then

$$\sup_{\omega \in \Omega} \mathbb{E}d^2(\hat{\omega}, \omega) \geq \min_{i \neq j} \frac{d^2(\omega_i, \omega_j)}{4} \frac{\sqrt{m_*}}{1 + \sqrt{m_*}} \left(1 - 2c - \sqrt{\frac{2c}{\log m_*}}\right).$$

To apply the Fano's Lemma, we need to find a collection of least favorable parameters $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{m_*}\}$ such that the difficulty of estimation among this subclass is almost the same as that among the whole sparsity class $\mathcal{F}_q^{p_1, p_2}(s_1, s_2, C_\lambda)$. To be specific, we check that the distance $d^2(\omega_i, \omega_j)$ among this collection of least favorable parameters is lower bounded by the sharp rate of the convergence and the average Kullback-Leibler divergence is indeed bounded above by the logarithm cardinality of the collection, i.e. Equation (32). In the proof we will show this via three main steps. Before that, we need two auxiliary lemmas.

LEMMA G.2. *Let* $\{0, 1\}^{p_1 - 1}$ *be equipped with Hamming distance* δ. *For integer* $0 < d < \frac{p_1 - 1}{4}$, *there exists some subset* $\Phi = \{\phi_1, \ldots, \phi_m\} \subset \{0, 1\}^{p_1 - 1}$ *such that*

$$\delta(\phi_i, \phi_j) \geq \frac{d}{2} \quad \text{for any} \quad \phi_i \neq \phi_j, \tag{33}$$

$$\delta\left(\phi_i, \vec{0}\right) = d \quad \text{for any} \quad \phi_i, \tag{34}$$

$$\log m \geq C_0 d \log\left(\frac{p_1}{d}\right) \text{for some constant } C_0. \tag{35}$$

See Massart (2007), Lemma 4.10 for more details.

LEMMA G.3. *For* $i = 1, 2$, *let* θ_i *and* η_i *be some unit vectors and* $\mathbb{P}_i$ *be the distribution of* n *i.i.d.* $N(0, \Sigma_i)$, *where the covariance matrix is defined as*

$$\Sigma_i = \begin{pmatrix} I_{p_1 \times p_1} & \lambda \theta_i \eta_i^T \\ \lambda \eta_i \theta_i^T & I_{p_2 \times p_2} \end{pmatrix}.$$

Then we have

$$K(\mathbb{P}_1, \mathbb{P}_2) \leq \frac{n\lambda^2}{2(1 - \lambda^2)} \left(\|\theta_1 - \theta_2\|^2 + \|\eta_1 - \eta_2\|^2\right).$$

We shall divide the proof into three main steps.

Step 1: Constructing the parameter set. Without loss of generality we assume $C_\lambda < \frac{1}{2}$. The subclass of parameters we will pick can be described in the following form:

$$\Sigma_{(p_1 + p_2) \times (p_1 + p_2)} = \begin{pmatrix} I_{p_1 \times p_1} & \frac{1}{2}\theta\eta^T \\ \frac{1}{2}\eta\theta^T & I_{p_2 \times p_2} \end{pmatrix},$$

where we will pick a collection of θ or η such that they are separated with the right rate of convergence. Without loss of generality, we assume that $s_1 \geq s_2$ and hence $s = s_1$. If

the inequality $s_1 \geq s_2$ holds in the other direction, we only need to switch the roles of θ and η. In this case, we will pick the collection of least favorable parameters indexed by the canonical pair $\omega = (\theta, \eta)$. Specifically, define $\Omega = \{\omega_0, \omega_1, \ldots, \omega_{m_*}\}$ where $\omega_i = (\theta_i, e_1)$ and e_1 is the unit vector in $\mathbb{R}^{p_2}$ with the first coordinate 1 and all others 0. The number m_* will be determined by the a version of Varshamov-Gilbert bound in Lemma G.2.

Now we define $m_* = m - 1$ and each $\theta_{i-1} = \left(\left(1 - \epsilon^2\right)^{1/2}, \phi_i \epsilon d^{-1/2} \right)$, where we pick

$$\epsilon = c_1 (s_1 - 1)^{1/2} \left(\frac{\log p_1}{n} \right)^{\frac{1}{2} - \frac{q}{4}} \text{ and } d = (s_1 - 1) \left(\frac{\log p_1}{n} \right)^{-\frac{q}{2}},$$

while ϕ_i and m are determined in Lemma G.2 accordingly. The constant $c_1 \in (0, 1)$ is to be determined later. It's easy to check that each θ_{i-1} is a unit vector. By our sparsity assumption $1 - \epsilon^2 \geq 0.5$ by picking a sufficient small constant c_1 and consequently the first coordinate is the largest one in magnitude. Clearly $\left|\theta_{i,(1)}\right|^q \leq s_1$. Moreover, we have

$$\begin{aligned} \left|\theta_{i,(k)}\right|^q &= \epsilon^q d^{-q/2} = c_1^q \left(\frac{\log p_1}{n} \right)^{\frac{q}{2}} \leq s_1 k^{-1}, 2 \leq k \leq d + 1, \\ \left|\theta_{i,(k)}\right|^q &= 0 \leq s_1 k^{-1}, k > d + 1. \end{aligned}$$

Hence each θ_i is in the corresponding weak l_q ball. Therefore our parameter subclass $\Omega \subset \mathcal{F}_q^{p_1, p_2} (s_1, s_2, C_\lambda)$.

Step 2: Bounding $d^2 (\omega_i, \omega_j)$. The loss function we considered in this section for ω_i and ω_j can be simplified as

$$d^2 (\omega_i, \omega_j) = L^2(\theta_i, \theta_j) = \left\| \theta_i \theta_i^T - \theta_j \theta_j^T \right\|_F^2 \geq \left\| \theta_j - \theta_i \right\|^2,$$

whenever $\|\theta_j - \theta_i\|^2 \leq 2$ which is satisfied in our setting since $\|\theta_j - \theta_i\|^2 \leq 2\epsilon^2 \leq 1$. The Equation (33) in Lemma G.2 implies that

$$\begin{aligned} \min_{i \neq j} d^2 (\omega_i, \omega_j) &\geq \frac{d}{2} \left(\epsilon d^{-1/2} \right)^2 \\ &\geq \frac{\epsilon^2}{2} = \frac{c_1^2}{2} (s_1 - 1) \left(\frac{\log p_1}{n} \right)^{1 - \frac{q}{2}}. \end{aligned} \tag{36}$$

which is the sharp rate of convergence, noting that the Equation (36) is still true up to a constant when we replace p_1 by p and s_1 by s.

Step 3: Bounding the Kullback-Leibler divergence.

Note that in our case $\eta_1 = \eta_2 = e_1$. Lemma G.3, together with Equations (34) and (35), imply

$$
\begin{aligned}
\frac{1}{m_*} \sum_{1 \le k \le m_*} K\left(\mathbb{P}_{\omega_k}, \mathbb{P}_{\omega_0}\right) &\le \frac{n\lambda^2}{2\left(1 - \lambda^2\right)} \frac{d}{2} \left(\epsilon d^{-1/2}\right)^2 \\
&= \frac{n\lambda^2}{4\left(1 - \lambda^2\right)} c_1^2 \left(s_1 - 1\right) \left(\frac{\log p_1}{n}\right)^{1 - \frac{q}{2}} \\
&= \frac{n}{12} c_1^2 d \frac{\log p_1}{n} < \frac{1}{10} \log m,
\end{aligned}
$$

where the last inequality is followed by picking a sufficiently small constant $c_1 > 0$ and noting that $\lambda = \frac{1}{2}$, $s_i \left(\frac{n}{\log p_i}\right)^{q/2} = o(p_i)$ in the assumption. Therefore we could apply Lemma G.1 and Equation (36) to obtain the sharp rate of convergence, which completes our proof.

$$
\sup_{P \in \mathcal{F}} \mathbb{E}_P \left(L^2(\hat{\theta}, \theta) \vee L^2(\hat{\eta}, \eta)\right) \ge Cs \left(\frac{\log p}{n}\right)^{1 - \frac{q}{2}},
$$

for any estimator $(\hat{\theta}, \hat{\eta})$, where $\mathcal{F} = \mathcal{F}_q^{p_1, p_2}\left(s_1, s_2, C_\lambda\right)$. Hence, Theorem 4.2 is proved.

We finally prove Lemma G.3 to complete the whole proof.

PROOF (PROOF OF LEMMA G.3). Let's rewrite matrix Σ_i in the following way

$$
\Sigma_i = I + \frac{\lambda}{2} \begin{pmatrix} \theta_i \\ \eta_i \end{pmatrix} \begin{pmatrix} \theta_i \\ \eta_i \end{pmatrix}^T - \frac{\lambda}{2} \begin{pmatrix} \theta_i \\ -\eta_i \end{pmatrix} \begin{pmatrix} \theta_i \\ -\eta_i \end{pmatrix}^T.
$$

Note that $\left\|\left(\theta_i^T, \eta_i^T\right)\right\| = \sqrt{2}$, so Σ_1 and Σ_2 have the same eigenvalues $1 + \lambda, 1, 1, \ldots, 1 - \lambda$. Thus we have

$$
\begin{aligned}
K\left(\mathbb{P}_1, \mathbb{P}_2\right) &= \frac{n}{2} \left[\operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right) - p - \log \det\left(\Sigma_2^{-1}\Sigma_1\right)\right] \\
&= \frac{n}{2} \left[\operatorname{tr}\left(\Sigma_2^{-1}\Sigma_1\right) - p\right] = \frac{n}{2} \operatorname{tr}\left(\Sigma_2^{-1}\left(\Sigma_1 - \Sigma_2\right)\right) \\
&= \frac{n\lambda}{2} \operatorname{tr}\left(\Sigma_2^{-1} \begin{pmatrix} 0 & A_1 - A_2 \\ A_1^T - A_2^T & 0 \end{pmatrix}\right),
\end{aligned}
\tag{37}
$$

where $p = p_1 + p_2$ and $A_i = \theta_i \eta_i^T$. To explicitly write down the inverse of Σ_2, we use its eigen-decomposition

$$
\Sigma_2 = I - \begin{pmatrix} \theta_2\theta_2^T & 0 \\ 0 & \eta_2\eta_2^T \end{pmatrix} + \frac{1 + \lambda}{2} \begin{pmatrix} \theta_2 \\ \eta_2 \end{pmatrix} \begin{pmatrix} \theta_2 \\ \eta_2 \end{pmatrix}^T + \frac{1 - \lambda}{2} \begin{pmatrix} \theta_2 \\ -\eta_2 \end{pmatrix} \begin{pmatrix} \theta_2 \\ -\eta_2 \end{pmatrix}^T.
$$

Thus we have

$$
\begin{aligned}
\Sigma_2^{-1} &= I - \begin{pmatrix} \theta_2\theta_2^T & 0 \\ 0 & \eta_2\eta_2^T \end{pmatrix} + \frac{1}{2\left(1 + \lambda\right)} \begin{pmatrix} \theta_2 \\ \eta_2 \end{pmatrix} \begin{pmatrix} \theta_2 \\ \eta_2 \end{pmatrix}^T + \frac{1}{2\left(1 - \lambda\right)} \begin{pmatrix} \theta_2 \\ -\eta_2 \end{pmatrix} \begin{pmatrix} \theta_2 \\ -\eta_2 \end{pmatrix}^T, \\
&= I - \frac{\lambda}{2\left(1 + \lambda\right)} \begin{pmatrix} \theta_2 \\ \eta_2 \end{pmatrix} \begin{pmatrix} \theta_2 \\ \eta_2 \end{pmatrix}^T + \frac{\lambda}{2\left(1 - \lambda\right)} \begin{pmatrix} \theta_2 \\ -\eta_2 \end{pmatrix} \begin{pmatrix} \theta_2 \\ -\eta_2 \end{pmatrix}^T
\end{aligned}
$$

Plugging this representation into Equation (37), we obtain that

$$
\begin{aligned}
K\left(\mathbb{P}_1, \mathbb{P}_2\right) &= \frac{n\lambda^2}{2}\left[\operatorname{tr}\left(\frac{A_2\left(A_2^T - A_1^T\right)}{1-\lambda^2}\right) + \operatorname{tr}\left(\frac{A_2^T\left(A_2 - A_1\right)}{1-\lambda^2}\right)\right] \\
&= \frac{n\lambda^2}{1-\lambda^2}\operatorname{tr}\left(A_2\left(A_2^T - A_1^T\right)\right) \\
&= \frac{n\lambda^2}{1-\lambda^2}\left(1 - \left(\theta_1^T\theta_2\right)\left(\eta_1^T\eta_2\right)\right) \\
&= \frac{n\lambda^2}{1-\lambda^2}\left(1 - \left(1 - \frac{\|\theta_1 - \theta_2\|^2}{2}\right)\left(1 - \frac{\|\eta_1 - \eta_2\|^2}{2}\right)\right) \\
&\leq \frac{n\lambda^2}{2\left(1-\lambda^2\right)}\left(\|\theta_1 - \theta_2\|^2 + \|\eta_1 - \eta_2\|^2\right),
\end{aligned}
$$

where we used that θ_i and η_i are unit vectors in the fourth equation above.

References

Avants, B. B., P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman (2010). Dementia induces correlated reductions in white matter integrity and cortical thickness: a multivariate neuroimaging study with sparse canonical correlation analysis. *Neuroimage 50*(3), 1004–1016.

Bach, F. R. and M. I. Jordan (2005). A probabilistic interpretation of canonical correlation analysis. *preprint*.

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: methodological and philosophical foundations. *MIS Quarterly 35*(2), 261–292.

Bickel, P. J. and E. Levina (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227.

Bickel, P. J. and E. Levina (2008b). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Bin, G., X. Gao, Z. Yan, B. Hong, and S. Gao (2009). An online multi-channel ssvep-based brain–computer interface using a canonical correlation analysis method. *Journal of Neural Engineering 6*(4), 046002.

Birnbaum, A., I. M. Johnstone, B. Nadler, and D. Paul (2012). Minimax bounds for sparse pca with noisy high-dimensional data. *arXiv preprint arXiv:1203.0967*.

Cai, T., W. Liu, and X. Luo (2011). A constrained 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494).

Cai, T. T., Z. Ma, and Y. Wu (2012). Sparse pca: Optimal rates and adaptive estimation. *arXiv preprint arXiv:1211.1309*.

Cai, T. T., Z. Ren, and H. H. Zhou (2013). Optimal rates of convergence for estimating toeplitz covariance matrices. *Probability Theory and Related Fields 156*(1-2), 101–143.

Cai, T. T. and M. Yuan (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics 40*(4), 2014–2042.

Cai, T. T., C. H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics 38*(4), 2118–2144.

Cai, T. T. and H. H. Zhou (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics 40*(5), 2389–2420.

Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. Kibbe, L. Hou, and S. Lin (2010). Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics 11*(1), 587.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Golub, G. H. and C. F. Van Loan (1996). Matrix computations.

Hardoon, D. R. and J. Shawe-Taylor (2011). Sparse canonical correlation analysis. *Machine Learning 83*(3), 331–353.

Hardoon, D. R., S. Szedmak, and J. Shawe-Taylor (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation 16*(12), 2639–2664.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika 28*(3/4), 321–377.

Johnstone, I. (2001). Thresholding for weighted chi-squared. *Statistica Sinica 11*(3), 691–704.

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104* (486).

Lê Cao, K.-A., P. G. Martin, C. Robert-Granié, and P. Besse (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics 10* (1), 34.

Ma, X.-J., R. Salunga, J. T. Tuggle, J. Gaudet, E. Enright, P. McQuary, T. Payette, M. Pistone, K. Stecker, B. M. Zhang, et al. (2003). Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences 100* (10), 5974–5979.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41* (2), 772–801.

Massart, P. (2007). Concentration inequalities and model selection.

Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34* (3), 1436–1462.

Parkhomenko, E., D. Tritchler, and J. Beyene (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology 8* (1), 1–34.

Ren, Z., T. Sun, C.-H. Zhang, and H. H. Zhou (2013). Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*.

Stewart, G. W. and J.-g. Sun (1990). Matrix perturbation theory.

TCGA (2012). Comprehensive molecular portraits of human breast tumours. *Nature 490*, 61–70.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61* (3), 611–622.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

VanderKraats, N. D., J. F. Hiken, K. F. Decker, and J. R. Edwards (2013). Discovering high-resolution patterns of differential dna methylation that correlate with gene expression changes. *Nucleic Acids Research*.

van't Veer, L. J., H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415*(6871), 530–536.

Waaijenborg, S. and A. H. Zwinderman (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics 10*(1), 315.

Weigelt, B., Z. Hu, X. He, C. Livasy, L. A. Carey, M. G. Ewend, A. M. Glas, C. M. Perou, and L. J. van't Veer (2005). Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Research 65*(20), 9155–9158.

Wiesel, A., M. Kliger, and A. O. Hero III (2008). A greedy approach to sparse canonical correlation analysis. *arXiv preprint arXiv:0801.2748*.

Witten, D., R. Tibshirani, S. Gross, B. Narasimhan, and M. D. Witten (2013). Package pma. *Genetics and Molecular Biology 8*(1), 28.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Witten, D. M. and R. J. Tibshirani (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology 8*(1), 1–27.

Yang, D., Z. Ma, and A. Buja (2013). A sparse svd method for high-dimensional data. *arXiv preprint arXiv:1112.2433*.