An Efficient and Optimal Method for Sparse Canonical Correlation Analysis

Chao Gao[1], Zongming Ma[2], and Harrison H. Zhou[1]

[1] *Yale University*

[2] *University of Pennsylvania*

Abstract

Canonical correlation analysis (CCA) is an important multivariate technique for exploring the relationship between two sets of variables which finds applications in many fields. This paper considers the problem of estimating the subspaces spanned by sparse leading canonical correlation directions when the ambient dimensions are high. We propose a computationally efficient two-stage estimation procedure which consists of a convex programming based initialization stage and a group Lasso based refinement stage. Moreover, we show that our procedure achieves optimal rates of convergence under mild conditions by deriving both the error bounds of the proposed estimator and the matching minimax lower bounds. In particular, the computation of the estimator does not involve estimating the marginal covariance matrices of the two sets of variables, and its minimax rate optimality requires no structural assumption on the marginal covariance matrices as long as they are well conditioned. The procedure yields encouraging numerical results on simulated datasets, and its practical usefulness is demonstrated by an application on a breast cancer dataset.

Keywords. Convex programming, Group Lasso, Minimax rates, Rates of convergence, Sparse CCA (SCCA)

1 Introduction

Canonical correlation analysis (CCA) [14] is one of the most classical and important tools in multivariate statistics [1, 20]. For two centered random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$, CCA finds matrices $U \in \mathbb{R}^{p\times r}$ and $V \in \mathbb{R}^{m\times r}$, such that the correlation between the two low dimensional vectors $U'X$ and $V'Y$ are maximized. To be precise, (U, V) solves to the following program,

$$\text{maximize } \mathsf{Tr}(L'\Sigma_{xy}R), \quad \text{subject to} \quad L'\Sigma_x L = R'\Sigma_y R = I_r, \tag{1}$$

where $\Sigma_x = \mathbb{E}XX', \Sigma_y = \mathbb{E}YY'$, and $\Sigma_{xy} = \mathbb{E}XY'$. Such technique is widely used in various scientific fields to explore the relation between two sets of variables.

Recently, there is a growing interest in applying CCA to high dimensional data analysis, where the dimensions p and m could be much larger than the sample size n. In such regime, the classical CCA does not work because the singular value decomposition method by [14] requires the invertibility of the marginal sample covariance matrices, which is not true when $p \vee m > n$. Motivated by genomics, neuroimaging and other applications, sparsity assumptions are imposed on the leading canonical correlation directions. This is called sparse canonical correlation analysis (SCCA), and various estimation procedures for SCCA have been developed in the literature. See, for example, [30, 31, 22, 13, 16, 27, 3, 28] for some recent methodological developments and applications.

In addition to progress on methodology, the theoretical aspect of SCCA has also been investigated in the literature. [9] showed that the (U, V) pair that solves (1) can be identified as population parameter if we rewrite Σ_{xy} as

$$\Sigma_{xy} = \Sigma_x (U \Lambda V' + U_1 \Lambda_1 V_1') \Sigma_y, \tag{2}$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_r)$, $\Lambda_1 = \text{diag}(\lambda_{r+1}, ..., \lambda_{p \wedge m})$ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{p \wedge m}$, and the constraints $U' \Sigma_x U = V' \Sigma_y V = I_r$, $U_1' \Sigma_x U_1 = V_1' \Sigma_y V_1 = I_{p \wedge m - r}$, $U' \Sigma_x U_1 = V' \Sigma_y V_1 = 0$ are satisfied. When $\Lambda_1 = 0$, this is called "multiple canonical pair" model, and in this case, the cross-covariance Σ_{xy} has a low-rank structure. Let $S_u = \text{supp}(U)$ and $S_v = \text{supp}(V)$ be the indices of nonzero rows of U and V. Then, SCCA means S_u and S_v have small cardinality. That is,

$$|S_u| \leq s_u \quad \text{and} \quad |S_v| \leq s_v. \tag{3}$$

Under this model, in a recent work, [11] showed that the minimax rate for SCCA under the loss function $\|\widehat{U}\widehat{V}' - UV'\|_{\text{F}}^2$ is

$$\frac{1}{n \lambda_r^2} \left(r(s_u + s_v) + s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v} \right), \tag{4}$$

under the assumption that $\lambda_{r+1} \leq c \lambda_r$ for some sufficiently small $c \in (0,1)$ and some mild regularity conditions. However, in [11], the upper bound is achieved by a computationally infeasible and nonadaptive procedure, which requires the knowledge of s_u and s_v and exhaustive search of all possible subsets with the given cardinality. In this paper, we raise a fundamental question: Is there a computationally efficient and sparsity-adaptive method which can achieve the optimal rate?

We provide an affirmative answer to this question under the multiple canonical pair model by proposing a two-stage estimation procedure called CoLaR, standing for Convex programming with Lasso Refinement. In the first stage, we propose a convex programming for SCCA based on a convex relaxation of a combinatorial program studied in [11]. The convex programming can be efficiently solved by the Alternating Direction Method with Multipliers (ADMM) algorithm [7]. Based on the output of the first stage, we formulate a sparse linear regression problem in the second stage to improve the rate of convergence, and the final estimator $\widehat{U}$ and $\widehat{V}$ can be obtained via a group-Lasso algorithm [33]. Under mild assumptions, we show that $\widehat{U}$ and $\widehat{V}$ recover the column spaces of U and V with the desired rate of convergence in probability. To be precise, for any matrix F, let P_F denote the projection matrix onto its column space. We show that for some constant $C > 0$,

$$\begin{aligned} \|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \leq & \quad C\frac{s_u(r + \log p)}{n\lambda_r^2}, \\ \|P_{\widehat{V}} - P_V\|_{\mathrm{F}}^2 \leq & \quad C\frac{s_v(r + \log m)}{n\lambda_r^2}, \end{aligned} \tag{5}$$

with high probability. The rate (5) is comparable to the minimax rate (4). To show (5) is optimal, we provide a minimax lower bound for the subspace loss in Section 3.2.

The foregoing result gives new insights on the problem of SCCA. To the best of our limited knowledge, [9] developed the first computationally efficient SCCA method which can provably achieve minimax optimal rates. They considered the special case of $r = 1$ and proposed an iterative thresholding method for estimating the sparse canonical directions. However, their estimation procedure requires the structural knowledge of the marginal inverse covariance matrices Σ_x^{-1} and Σ_y^{-1} and only achieves the optimal rates of convergence when the estimation errors of Σ_x^{-1} and Σ_y^{-1} are dominated by those of estimating the canonical directions. It is challenging to estimate Σ_x^{-1} and Σ_y^{-1} well in a high-dimensional setting. On the other hand, [11] showed the minimax rates of SCCA does not depend on the marginal covariance matrices as long as they are well-conditioned, though the upper bounds were achieved by a combinatorial programming which is computationally intractable. The result in the current paper complements that in [11] and shows that, even with no structural knowledge about the marginal covariance matrices, one can still obtain minimax rate optimal and sparsity-adaptive estimators via computationally efficient algorithms for a wide range of parameter spaces of interest.

Connection to the literature The current paper is related to the recent development on the sparse principal component analysis (SPCA) problem. For PCA, most literatures assume a spiked covariance structure where $\Sigma = V'\Lambda V + I_p$, with $V'V = I_r$, $\Lambda = (\lambda_1, ..., \lambda_r)$ and $\lambda_1 \geq ... \geq \lambda_r$. Conceptually, this is analogous to the multiple canonical pair model for SCCA considered in the current paper. For SPCA, Johnstone and Lu [15] proposed a diagonal thresholding estimator of the sparse principal eigenvector which is provably consistent when $r = 1$ in the spiked covariance model. A semidefinite relaxation of SPCA was proposed by [10], and was extended to the multiple-rank case by [26] using the fantope projection idea. An iterative thresholding scheme was developed by [19] for principal subspace estimation. A regression formulation of SPCA was proposed in [8].

Organization The rest of the paper is organized as follows. In Section 2, we propose our estimation procedure. Its statistical optimality is analyzed in Section 3, where we present rates of convergence and the corresponding minimax lower bounds. In Section 4 we demonstrate the competitive finite sample performance of our approach by numerical experiments. The proofs of the main results are presented in Section 6 with some additional technical details deferred to Section 7.

Notation For a positive integer t, $[t]$ denotes the index set $\{1, 2, ..., t\}$. For any set S, $|S|$ denotes its cardinality. For any event E, $\mathbf{1}_{\{E\}}$ denotes its indicator function. For any number a, we use $\lceil a \rceil$ to denote the smallest integer that is no smaller than a. For any two numbers a and b, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For a vector u, $||u|| = \sqrt{\sum_i u_i^2}$, $||u||_0 = \sum_i \mathbf{1}_{\{u_i \neq 0\}}$ and $||u||_1 = \sum_i |u_i|$. For any matrix $A = (a_{ij})_{i\in[p], j\in[k]}$, the i-th row of A is denoted by $A_{i\cdot}$. For subsets $J \subset [p] \times [k]$ of indices, we use notation $A_J = (a_{ij}\mathbf{1}_{\{(i,j)\in J\}})$. When $J = J_1 \times J_2$ with $J_1 \subset [p]$ and $J_2 \subset [k]$, we write $A_{J_1 J_2}$ to stand for $A_{J_1 \times J_2}$ and write $A_{(J_1 J_2)^c}$ to stand for $A_{(J_1 \times J_2)^c}$.

The notation $A_{J_1 *}$ means $A_{J_1 \times [k]}$. For any square matrix $A = (a_{ij})$, denote its trace by $\mathsf{Tr}(A) = \sum_i a_{ii}$. For two square matrices A, B, the relation $A \preceq B$ means $B - A$ is positive semidefinite. Moreover, let $O(p, k)$ denote the set of all $p \times k$ orthonormal matrices and $O(k) = O(k, k)$. For any matrix $A \in \mathbb{R}^{p \times k}$, $\sigma_i(A)$ stands for its i-th largest singular value. The Frobenius norm and the operator norm of A are defined as $\|A\|_{\mathrm{F}} = \sqrt{\mathsf{Tr}(A'A)}$ and $\|A\|_{\mathrm{op}} = \sigma_1(A)$, respectively. The l_1 norm and the nuclear norm are defined as $||A||_1 = \sum_{ij} |a_{ij}|$ and $\|A\|_* = \sum_i \sigma_i(A)$, respectively. The support of A is defined as $\mathrm{supp}(A) = \{i \in [p] : \|A_{i \cdot}\| > 0\}$. For any positive semi-definite matrix A, $A^{1/2}$ denotes its principal square root that is positive semi-definite and satisfies $A^{1/2}A^{1/2} = A$. The trace inner product of two matrices $A, B \in \mathbb{R}^{p \times k}$ is defined as $\langle A, B \rangle = \mathsf{Tr}(A'B)$. The constant C and its variants such as C_1, C', etc. are generic constants and may vary from line to line, unless otherwise specified.

2 Methodology

In this section, we introduce our methodology, CoLaR, for estimating the canonical correlation matrices U and V. The estimation procedure is divided into two stages: initialization and refinement. They are detailed out in Sections 2.1 and 2.2, respectively.

2.1 Initialization by Convex Programming

Suppose we have i.i.d. observations $(X_i, Y_i)_{i=1}^n$ from some centered distribution, where $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}^m$, and let

$$\widehat{\Sigma} = \begin{bmatrix} \widehat{\Sigma}_x & \widehat{\Sigma}_{xy} \\ \widehat{\Sigma}_{yx} & \widehat{\Sigma}_y \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} X_i \\ Y_i \end{bmatrix} \begin{bmatrix} X_i' & Y_i' \end{bmatrix}$$

be the joint sample covariance matrix. [11] showed that the solution to the following optimization problem is optimal for sparse CCA:

$$\begin{aligned} &\text{maximize} \quad \mathsf{Tr}(L'\widehat{\Sigma}_{xy}R) \\ &\text{subject to} \quad L'\widehat{\Sigma}_x L = R'\widehat{\Sigma}_y R = I_r \;\text{ and }\; |\text{supp}(L)| = s_u, |\text{supp}(R)| = s_v. \end{aligned} \tag{6}$$

However, (6) is not computationally feasible because solving it requires searching over all s_u subsets of $[p]$ and s_v subsets of $[m]$, and the computational cost grows exponentially with the dimension of the problem. Moreover, (6) depends on the true sparsity s_u and s_v, and it is not adaptive, either.

This motivates us to consider the following convex relaxation of the program (6). First, note that the objective can be written as

$$\mathsf{Tr}(L'\widehat{\Sigma}_{xy}R) = \left\langle \widehat{\Sigma}_{xy}, LR' \right\rangle.$$

Thus, it is linear with respect to LR'. This suggests to treat the matrix LR' as a single object instead of optimizing over L and R separately. The constraints $|\text{supp}(L)| = s_u, |\text{supp}(R)| = s_v$ implies that LR' has at most $s_u s_v$ nonzero entries. Relaxing the l_0 norm by the l_1 norm, we obtain the new objective function

$$\left\langle \widehat{\Sigma}_{xy}, F \right\rangle - \rho ||F||_1, \tag{7}$$

where F serves as a surrogate for LR'. To deal with the other constraints $L'\widehat{\Sigma}_x L = R'\widehat{\Sigma}_y R = I_r$, note that they are equivalent to $\widehat{\Sigma}_x^{1/2} L \in O(p, r)$ and $\widehat{\Sigma}_y^{1/2} R \in O(m, r)$. Let $G = \widehat{\Sigma}_x^{1/2} LR' \widehat{\Sigma}_y^{1/2} = \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2}$. Since it is a product of two orthogonal matrices, its operator norm is bounded by 1. Together with the fact that its rank is not more than r, the nuclear norm is also bounded by r. Thus, it belongs to the following convex set

$$\{G : \|G\|_* \le r, \|G\|_{\text{op}} \le 1\}. \tag{8}$$

Combining (7) and (8), we obtain the following convex relaxation of (6),

$$\begin{aligned} &\text{maximize} \quad \left\langle \widehat{\Sigma}_{xy}, F \right\rangle - \rho||F||_1, \\ &\text{subject to} \quad \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2}\|_* \leq r, \\ &\qquad\qquad\quad \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2}\|_{\text{op}} \leq 1. \end{aligned} \tag{9}$$

Intuitively, the solution to (9) should be a good estimator for UV'.

A similar convex relaxation was proposed by [26] for sparse PCA. However, they constrained the projection matrix onto the span of the leading eigenvectors to the fantope $\{P : \mathsf{Tr}(P) = r, 0 \preceq P \preceq I_p\}$, which is a convex set of P. Note that such a relaxation is not directly applicable here, since for the projection matrix $P = GG' = \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y F' \widehat{\Sigma}_x^{1/2}$ of interest, the constraint $\mathsf{Tr}(P) = \mathsf{Tr}(\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y F' \widehat{\Sigma}_x^{1/2}) = r$ is not convex in F. The same is true for the projection matrix $G'G$. We propose a new relaxation (8) to overcome this issue in the sparse CCA problem.

2.1.1 Implementation via ADMM

To implement the convex programming (9), we turn to the Alternating Direction Method of Multipliers (ADMM) [7]. Note that (9) can be rewritten as

$$\begin{aligned} &\text{minimize} \quad f(F) + g(G), \\ &\text{subject to} \quad \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G = 0, \end{aligned} \tag{10}$$

where

$$f(F) = -\left\langle \widehat{\Sigma}_{xy}, F \right\rangle + \rho\|F\|_1, \tag{11}$$

$$g(G) = \infty \mathbf{1}_{\{\|G\|_* > r\}} + \infty \mathbf{1}_{\{\|G\|_{\text{op}} > 1\}}. \tag{12}$$

Thus, the augmented Lagrangian form of the problem is

$$\mathcal{L}_\eta(F, G, H) = f(F) + g(G) + \left\langle H, \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G \right\rangle + \frac{\eta}{2}\|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G\|_{\text{F}}^2. \tag{13}$$

Following the generic algorithm spelled out in Section 3 of [7], suppose after the k-th iteration, the matrices are (F^k, G^k, H^k), then we update the matrices in the $(k+1)$-th iteration as follows:

$$F^{k+1} = \underset{F}{\operatorname{argmin}} \mathcal{L}_\eta(F, G^k, H^k), \tag{14}$$

$$G^{k+1} = \underset{G}{\operatorname{argmin}} \mathcal{L}_\eta(F^{k+1}, G, H^k), \tag{15}$$

$$H^{k+1} = H^k + \eta(\widehat{\Sigma}_x^{1/2} F^{k+1} \widehat{\Sigma}_y^{1/2} - G^{k+1}). \tag{16}$$

The algorithm iterates over (14) – (16) till some convergence criterion is met. It is clear that the update (16) for the dual variable H is easy to calculate. Moreover the updates (14) and (15) can be solved easily and have explicit meaning in giving solution to SCCA. We are going to show that (14) is equivalent to a Lasso problem. Thus, this step targets at the sparsity of the matrix UV'. The update (15) turns out to be equivalent to a singular value capped soft thresholding problem, and it targets at the low-rankness of the matrix $\Sigma_x^{1/2} UV' \Sigma_y^{1/2}$. In what follows, we study in more details on the updates for F and G.

First, we note that (14) is equivalent to

$$\begin{aligned} F^{k+1} &= \underset{F}{\operatorname{argmin}} f(F) + \left\langle H^k, \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} \right\rangle + \frac{\eta}{2} \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G^k\|_{\mathrm{F}}^2 \\ &= \underset{F}{\operatorname{argmin}} \frac{\eta}{2} \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - (G^k - \frac{1}{\eta} H^k + \frac{1}{\eta} \widehat{\Sigma}_x^{-1/2} \widehat{\Sigma}_{xy} \widehat{\Sigma}_y^{-1/2})\|_{\mathrm{F}}^2 + \rho \|F\|_1. \end{aligned} \tag{17}$$

Thus, it is clear that the update of F in (14) reduces to a standard Lasso problem as summarized in the following proposition, which can be solved by standard software, such as TFOCS [4]. Here and after, for any positive semi-definite matrix A with eigen-decomposition $A = \sum_{i=1}^r \lambda_i \theta_i \theta_i'$ where r is the rank of A and the λ_i's are the nonzero eigenvalues with θ_i the associated eigenvectors, we define $A^{-1/2} = \sum_{i=1}^r \lambda_i^{-1/2} \theta_i \theta_i'$.

Proposition 2.1. *Let* vec *be the vectorization operation of any matrix and* $\otimes$ *the*

Kronecker product. Then $\mathrm{vec}(F^{k+1})$ *is the solution to the Lasso problem*

$$\text{minimize}_x \quad \|\Gamma x - b\|_{\mathrm{F}}^2 + \frac{2\rho}{\eta}\|b\|_1$$

where $\Gamma = \widehat{\Sigma}_y^{1/2} \otimes \widehat{\Sigma}_x^{1/2}$ *and* $b = \mathrm{vec}(G^k - \frac{1}{\eta}H^k + \frac{1}{\eta}\widehat{\Sigma}_x^{-1/2}\widehat{\Sigma}_{xy}\widehat{\Sigma}_y^{-1/2})$*.*

Since each update of F is the solution of some Lasso problem, it should be sparse in the sense that its vector l_1 norm is well controlled.

Turning to the update for G, we note that (15) is equivalent to

$$\begin{aligned}
G^{k+1} &= \operatorname*{argmin}_G g(G) - \langle H^k, G\rangle + \frac{\eta}{2}\|\widehat{\Sigma}_x^{1/2}F^{k+1}\widehat{\Sigma}_y^{1/2} - G\|_{\mathrm{F}}^2 \\
&= \operatorname*{argmin}_G \frac{\eta}{2}\|G - (\frac{1}{\eta}H^k + \widehat{\Sigma}_x^{1/2}F^{k+1}\widehat{\Sigma}_y^{1/2})\|_{\mathrm{F}}^2 \\
&\qquad\qquad + \infty\mathbf{1}_{\{\|G\|_* > r\}} + \infty\mathbf{1}_{\{\|G\|_{\mathrm{op}} > 1\}} \\
&= \operatorname*{argmin}_G \|G - (\frac{1}{\eta}H^k + \widehat{\Sigma}_x^{1/2}F^{k+1}\widehat{\Sigma}_y^{1/2})\|_{\mathrm{F}}^2 \\
&\qquad\qquad + \infty\mathbf{1}_{\{\|G\|_* > r\}} + \infty\mathbf{1}_{\{\|G\|_{\mathrm{op}} > 1\}}.
\end{aligned} \tag{18}$$

The solution to the last display has a closed form according to the following result.

Proposition 2.2. *Let* G^* *be the solution to the optimization problem:*

$$\begin{aligned}
&\text{minimize} \quad \|G - W\|_{\mathrm{F}} \\
&\text{subject to} \quad \|G\|_* \le r, \quad \|G\|_{\mathrm{op}} \le 1.
\end{aligned}$$

Let the SVD of W *be* $W = \sum_{i=1}^m \omega_i a_i b_i'$ *with* $\omega_1 \ge \cdots \ge \omega_m \ge 0$ *the ordered singular values. Then* $G^* = \sum_{i=1}^m g_i a_i b_i'$ *where for any* i*,* $g_i = 1 \wedge (\omega_i - \gamma^*)_+$ *for some* γ *which is the solution to*

$$\text{minimize} \quad \gamma, \qquad \text{subject to} \quad \gamma > 0, \quad \sum_{i=1}^m 1 \wedge (\omega_i - \gamma)_+ \le r.$$

Proof. The proof essentially follows that of Lemma 4.1 in [26]. In addition to the fact that the current problem deals with asymmetric matrix, the only difference that we now have an inequality constraint $\sum_i g_i \le r$ rather than an equality constraint as in

Algorithm 1: An ADMM algorithm for SCCA

Input:

1. Sample covariance matrices $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$ and $\widehat{\Sigma}_{xy}$,
2. Penalty parameter ρ,
3. Rank r,
4. ADMM parameter η and tolerance level ϵ.

Output: Estimated sparse canonical correlation signal $\widehat{A}$.

Initialize: $k = 0$, $F^0 = \mathrm{SVCST}(\widehat{\Sigma}_{xy}), G^0 = 0, H^0 = 0$.

repeat

  Update F^{k+1} as in (14) (Lasso) ;

  Update $G^{k+1} \leftarrow \mathrm{SVCST}(\eta^{-1}H^k + \widehat{\Sigma}_x^{1/2}F^{k+1}\widehat{\Sigma}_y^{1/2})$ (SVCST) ;

  Update $H^{k+1} \leftarrow H^k + \eta(\widehat{\Sigma}_x^{1/2}F^{k+1}\widehat{\Sigma}_y^{1/2} - G^{k+1})$;

  $k \leftarrow k + 1$;

until $\max\{\|F^{k+1} - F^k\|_{\mathrm{F}}, \rho\|G^{k+1} - G^k\|_{\mathrm{F}}\} \leq \epsilon$;

Return $\widehat{A} = F^k$.

[26]. The asymmetry of the current problem does not matter since it is orthogonally invariant. □

Here and after, we call the operation in Proposition 2.2 singular value capped soft thresholding (SVCST) and write $G^* = \mathrm{SVCST}(W)$. Thus, any update for G results from the SVCST operation of some matrix, and so it has well controlled singular values.

In summary, the convex program (9) is implemented as Algorithm 1.

2.2 Refinement by Sparse Regression

Let the optimizer be denoted by $\widehat{A}$. Let the columns of the matrix $U^{(0)}$ collect the first r left singular vectors of $\widehat{A}$ and the columns of $V^{(0)}$ collect the first r right singular vector of $\widehat{A}$. In Section 3, we show that the associated projection matrices $P_{U^{(0)}}$ and $P_{V^{(0)}}$ are consistent estimators of the subspace spanned by the columns of U and V respectively. The convergence rate is $\frac{s_u s_v \log(p+m)}{n\lambda_r^2}$ under the Frobenius loss. Compared with the minimax rate (4), it is sub-optimal. The reason is that the program (9) solves for $\widehat{A}$, which estimates UV', and the sparsity of the matrix UV' is $s_u s_v$ instead of s_u of U and s_v of V. To obtain the optimal convergence rate, we need a procedure directly estimating U and V.

To motivate such procedure, let us introduce a basic fact of CCA. Let (X, Y) have the same distribution as (X_i, Y_i). Consider the following least square problems

$$\min_{L\in\mathbb{R}^{p\times r}} \mathbb{E}\|L'X - V'Y\|_{\mathrm{F}}^2, \quad \min_{R\in\mathbb{R}^{m\times r}} \mathbb{E}\|R'Y - U'X\|_{\mathrm{F}}^2.$$

The solution is characterized by the following proposition.

Proposition 2.3. *Under the CCA structure (2), the minimizers of the above least square problems are $U\Lambda$ and $V\Lambda$.*

Proof. Since the objective is convex, the optimal L is achieved by equating the gradient to zero, which leads to $\Sigma_x L = \Sigma_{xy} V$. By (2), we have $\Sigma_{xy} V = \Sigma_x(U\Lambda V' + U_1\Lambda_1 V_1')\Sigma_y V = \Sigma_x U\Lambda$. Since Σ_x is invertible, we have $L = U\Lambda$. The same argument leads to $R = V\Lambda$. □

The result shows that if we have V, then we may find $U\Lambda$ by regressing $V'Y$ on X. On the other hand, if we have U, we can also find $V\Lambda$ by regressing $U'X$ on Y. With the estimator $U^{(0)}, V^{(0)}$ obtained from the convex programming (9), we propose the following sparse regression formulation of SCCA,

$$\begin{aligned} \widehat{U} &= \operatorname*{argmin}_{L\in\mathbb{R}^{p\times r}} \left\{ \frac{1}{n}\sum_{i=1}^n \|L'X_i - (V^{(0)})'Y_i\|_{\mathrm{F}}^2 + \rho_u \sum_{j=1}^p \|L_{j\cdot}\| \right\}, \\ \widehat{V} &= \operatorname*{argmin}_{R\in\mathbb{R}^{m\times r}} \left\{ \frac{1}{n}\sum_{i=1}^n \|R'Y_i - (U^{(0)})'X_i\|_{\mathrm{F}}^2 + \rho_v \sum_{j=1}^m \|R_{j\cdot}\| \right\}, \end{aligned} \quad (19)$$

where the penalty terms $\rho_u \sum_{j=1}^p \|L_{j\cdot}\|$ and $\rho_v \sum_{j=1}^m \|R_{j\cdot}\|$ encourages row-sparsity of $\widehat{U}$ and $\widehat{V}$. With simple algebra, (19) can be written only in terms of the sample covariance matrix $\widehat{\Sigma}$,

$$\begin{aligned} \widehat{U} &= \underset{L\in\mathbb{R}^{p\times r}}{\text{argmin}} \left\{ \mathsf{Tr}(L'\widehat{\Sigma}_x L) - 2\,\mathsf{Tr}(L'\widehat{\Sigma}_{xy}V^{(0)}) + \rho_u \sum_{j=1}^p ||L_{j\cdot}|| \right\}, \\ \widehat{V} &= \underset{R\in\mathbb{R}^{m\times r}}{\text{argmin}} \left\{ \mathsf{Tr}(R'\widehat{\Sigma}_y R) - 2\,\mathsf{Tr}(R'\widehat{\Sigma}_{yx}U^{(0)}) + \rho_v \sum_{j=1}^m ||R_{j\cdot}|| \right\}. \end{aligned} \tag{20}$$

The program (19) and its equivalent form (20) are essentially the group Lasso estimator proposed by [33], and it can be efficiently solved by standard software developed by [4]. We remark that it is critical to use the group Lasso penalty. If the naive l_1 penalty on the whole matrix is used, we will get a sub-optimal convergence rate.

3 Statistical Optimality

In this section, we show that the estimator proposed in Section 2 enjoys certain statistical optimality. The convergence rates of (9) and (20) are established in Section 3.1. A matching minimax lower bound is derived in Section 3.2. This shows that the estimator (20) initialized by (9) is minimax rate optimal.

3.1 Convergence Rates

In this section, we establish statistical properties of (9) and (20). We consider the multiple canonical pair model in [9], which corresponds to the CCA structure (2)-(3) with $\Lambda_1 = 0$. We define the parameter space $\mathcal{F}(p, m, s_u, s_v, r, \lambda_r; M)$ for the covariance by collecting all such Σ satisfying $\|\Sigma_x\|_{\text{op}} \vee \|\Sigma_y\|_{\text{op}} \vee \|\Sigma_x^{-1}\|_{\text{op}} \vee \|\Sigma_y^{-1}\|_{\text{op}} \leq M$ for some absolute constant $M > 0$. Define $Z \in \mathbb{R}^{p+m}$ as

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \Sigma^{1/2} Z, \tag{21}$$

and assume Z is an isotropic sub-Gaussian vector. To be precise, define the sub-Gaussian norm according to [24],

$$||Z||_{\psi_2} = \sup_{||b|| \leq 1} \inf \left\{ \xi > 0 : \mathbb{E} \exp \left| \frac{b'Z}{\xi} \right|^2 \leq 2 \right\}.$$

The class of distribution of the vector $(X', Y')'$ we consider is defined as

$$\begin{aligned} \mathcal{P}(p, m, s_u, s_v, r, \lambda_r; M) = \big\{ \mathbb{P} : (X', Y')' \sim \mathbb{P} \text{ has representation (21)}, \\ \text{with } \Sigma \in \mathcal{F}(p, m, s_u, s_v, r, \lambda_r; M), \\ \mathbb{E}Z = 0, ||Z||_{\psi_2} \leq 1 \big\}. \end{aligned}$$

In what follows, we also use $\mathbb{P}$ to implicitly represent the product measure $\mathbb{P}^n$.

For the program (9), recall that $U^{(0)}$ and $V^{(0)}$ are left and right singular vector matrices of rank r of the optimum $\widehat{A}$. The following theorem guarantees that the column spaces of $U^{(0)}$ and $V^{(0)}$ consistently recover the column spaces of U and V respectively.

Theorem 3.1. *Assume that*

$$\frac{s_u s_v \log(p+m)}{n\lambda_r^2} \leq c, \tag{22}$$

for some sufficiently small $c \in (0, 1)$. For any constant $C' > 0$, there exist constants $C > 0$ and $\gamma > 0$ only depending on M and C', such that when $\rho \geq \gamma\sqrt{\frac{\log(p+m)}{n}}$,

$$\|\widehat{A} - UV'\|_{\mathrm{F}}^2 \vee \|P_{U^{(0)}} - P_U\|_{\mathrm{F}}^2 \vee \|P_{V^{(0)}} - P_V\|_{\mathrm{F}}^2 \leq C \frac{s_u s_v \rho^2}{\lambda_r^2},$$

with $\mathbb{P}$-probability at least $1 - \exp\big(-C'(s_u + \log(ep/s_u))\big) - \exp\big(-C'(s_v + \log(em/s_v))\big)$ for any $\mathbb{P} \in \mathcal{P}(p, m, s_u, s_v, r, \lambda_r; M)$.

The program (20) uses $U^{(0)}$ and $V^{(0)}$ from the output of (9). It is possible to use other matrices. The following theorem guarantees the performance of (20) for an arbitrary $U^{(0)}, V^{(0)}$, which are independent of $\widehat{\Sigma}$.

Theorem 3.2. *Let $\widehat{U}, \widehat{V}$ be solutions to the program (20) initialized with matrices $U^{(0)}, V^{(0)}$ which are independent of $\widehat{\Sigma}$. Assume that*

$$\frac{r + \log p + \log m}{n} \leq C_1, \tag{23}$$

for some constant $C_1 > 0$. For any constant $C' > 0$, there exist positive constants γ_u, γ_v, C only depending on M, C' and C_1, such that when $\rho_u \geq \gamma_u \sqrt{\frac{r+\log p}{n}}$ and $\rho_v \geq \gamma_v \sqrt{\frac{r+\log m}{n}}$,

$$\begin{aligned} \|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 &\leq C \frac{s_u \rho_u^2}{\lambda_r^2 \sigma_{\min}^2 (V' \Sigma_y V^{(0)})}, \\ \|P_{\widehat{V}} - P_V\|_{\mathrm{F}}^2 &\leq C \frac{s_v \rho_v^2}{\lambda_r^2 \sigma_{\min}^2 (U' \Sigma_x U^{(0)})}, \end{aligned}$$

with $\mathbb{P}$-probability at least $1 - \exp\left(-C'(r + \log(p \wedge m))\right)$ for any $\mathbb{P} \in \mathcal{P}(p, m, s_u, s_v, r, \lambda_r; M)$.

Observe that $\Sigma_x^{1/2} U \in O(p, r)$ and $|\mathrm{supp}(U)| \leq s_u$ implicitly implies $r \leq s_u$, and similarly, $r \leq s_v$. Thus, the assumption (23) is implied by the assumption (22).

Note that as long as $\sigma_{\min}(V' \Sigma_y V^{(0)})$ and $\sigma_{\min}(U' \Sigma_x U^{(0)})$ are bounded away from zero, the rate of convergence of Theorem 3.2 is comparable to the minimax rate (4). This requires $(U^{(0)}, V^{(0)})$ being not too bad. Since Theorem 3.1 guarantees that $(U^{(0)}, V^{(0)})$ output from (9) has good statistical performance, we may combine (9) and (20). Let us split the sample into two halves, $\{(X_i, Y_i)\}_{i=1}^{\lceil n/2 \rceil}$ and $\{(X_i, Y_i)\}_{i=\lceil n/2 \rceil + 1}^{n}$. Let $(U^{(0)}, V^{(0)})$ be the output from (9) using $\{(X_i, Y_i)\}_{i=1}^{\lceil n/2 \rceil}$, and let $(\widehat{U}, \widehat{V})$ be the output of (20) using $\{(X_i, Y_i)\}_{i=1}^{\lceil n/2 \rceil}$ and initialized by $(U^{(0)}, V^{(0)})$. Then, we have the following result.

Theorem 3.3. *Assume (22). For any $C' > 0$, there exist constants γ, γ_u and γ_v depending only on c, C' and M such that if we set $\rho = \gamma' \sqrt{\frac{\log(p+m)}{n}}$, $\rho_u = \gamma_u' \sqrt{\frac{r+\log p}{n}}$ and $\rho_v = \gamma_v' \sqrt{\frac{r+\log m}{n}}$ for any $\gamma' \in [\gamma, C_2\gamma], \gamma_u' \in [\gamma_u, C_2\gamma_u]$ and $\gamma_v' \in [\gamma_v, C_2\gamma_v]$ for some absolute constant $C_2 > 0$, then there exists a constant $C > 0$ only depending on*

M, C', C_2 and c in (22), such that

$$\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \leq C\frac{s_u(r+\log p)}{n\lambda_r^2},$$
$$\|P_{\widehat{V}} - P_V\|_{\mathrm{F}}^2 \leq C\frac{s_v(r+\log m)}{n\lambda_r^2},$$

with $\mathbb{P}$-probability at least $1-\exp\left(-C'(s_u+\log(ep/s_u))\right)-\exp\left(-C'(s_v+\log(em/s_v))\right)-\exp\left(-C'(r+\log(p\wedge m))\right)$ for any $\mathbb{P}\in\mathcal{P}(p,m,s_u,s_v,r,\lambda_r;M)$.

Remark 3.1. *The rates $\frac{s_u(r+\log p)}{n\lambda_r^2}$ and $\frac{s_v(r+\log m)}{n\lambda_r^2}$ are optimal according to Theorem 3.4. The group Lasso penalty in (20) plays an important role. If we simply use a Lasso penalty, then we will obtain the rates $\frac{rs_u\log p}{n\lambda_r^2}$ and $\frac{rs_v\log m}{n\lambda_r^2}$, which is clearly sub-optimal.*

3.2 A Minimax Lower Bound

Note that the minimax rate (4) is for the loss function $\|\widehat{U}\widehat{V}'-UV'\|_{\mathrm{F}}^2$. It does not directly imply that the rate obtained in Theorem 3.3 is optimal. We derive a matching lower bound for the result in Theorem 3.3 under the desired projection loss.

Theorem 3.4. *Assume $r\leq\frac{s_u\wedge s_v}{2}$, and there exists some $\eta\in(0,1)$, such that $\lambda_r\leq 1-\eta$, $s_u\leq p^{1-\eta}$ and $s_v\leq m^{1-\eta}$. Then, there exist some constant $C>0$ only depending on M and η and an absolute constant $c_0>0$, such that for any $\widehat{U}$ and $\widehat{V}$, we have*

$$\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{P}\left(\|P_{\widehat{U}}-P_U\|_{\mathrm{F}}^2\geq C\frac{s_u(r+\log p)}{n\lambda_r^2}\wedge c_0\right)\geq 0.8,$$
$$\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{P}\left(\|P_{\widehat{V}}-P_V\|_{\mathrm{F}}^2\geq C\frac{s_v(r+\log m)}{n\lambda_r^2}\wedge c_0\right)\geq 0.8,$$

where $\mathcal{P}=\mathcal{P}(p,m,s_u,s_v,r,\lambda_r;M)$.

4 Numerical Results

In this section, we present numerical results that demonstrate the finite sample performance of the proposed sparse CCA method on synthetic datasets. We consider four simulation settings and focus on the multiple canonical pair case where $r>1$.

Implementation details In all numerical results reported in this section, we used penalty level $0.55 \times \sqrt{\log(p \vee m)/n}$ in the initialization stage, and set the ADMM parameter $\eta = 2$ and tolerance $\epsilon = 10^{-4}$. In the refinement stage, we used five-fold cross validation to select a common penalty parameter used in group Lasso. For $l = 1, \ldots, 5$, we use one fold of the data as the test sample $(X_{(l)}^{\text{test}}, Y_{(l)}^{\text{test}})$ and the other four folds as the training sample $(X_{(l)}^{\text{train}}, Y_{(l)}^{\text{train}})$. For a particular choice of the penalty parameter $\rho_u = \rho_v = \rho$, we apply the refinement algorithm on $(X_{(l)}^{\text{train}}, Y_{(l)}^{\text{train}})$ to obtain estimates $(\widehat{U}_{(l)}, \widehat{V}_{(l)})$. Then we compute the sum of canonical correlations between $X_{(l)}^{\text{test}}\widehat{U}_{(l)}$ and $Y_{(l)}^{\text{test}}\widehat{V}_{(l)}$ to obtain $\text{CV}(\rho)$. Among all the candidate penalty parameters, we select the ρ value such that $\text{CV}(\rho)$ is maximized. The candidate penalty values used in the simulation below are $\{0.5, 1, 1.5, 2\} \times \sqrt{(r + \log(p \vee m))/n}$. We use all the sample $\{(X_i, Y_i)\}_{i=1}^n$ in both stages of the estimation procedure.

To demonstrate the competitive performance of the proposed SCCA method, we compare it with the method proposed in [31] (denoted by PMA here and on). The PMA is defined via the following optimization problem

$$\text{maximize } u'\widehat{\Sigma}_{xy}v, \quad \text{subject to } ||u|| \leq 1, ||v|| \leq 1, ||u||_1 \leq c_1, ||v||_1 \leq c_2.$$

The solution gives the first canonical pair $\widehat{u}_1, \widehat{v}_1$. The the same procedure is repeated after $\widehat{\Sigma}_{xy}$ is replaced by $\widehat{\Sigma}_{xy} - (\widehat{u}_1'\widehat{\Sigma}_{xy}\widehat{v}_1)\widehat{u}_1\widehat{v}_1'$, and the solution is the second canonical pair $\widehat{u}_2, \widehat{v}_2$. This process is repeated until $\widehat{u}_r, \widehat{v}_r$ is obtained. Note that the normalization constraint $\|u\| \leq 1$ and $\|v\| \leq 1$ implicitly assumes that the marginal covariance matrices Σ_x and Σ_y are identity matrices.

We used the R implementation of the method (PMA package in R) by the authors of [31] and the penalty parameter is always selected by cross validation by using the default settings.

Simulation settings In all four settings, we set $p = m$ and $\Sigma_x = \Sigma_y = \Sigma$ and $r = 2$ with $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$. Moreover, the nonzero rows of both U and V at

$\{1, 6, 11, 16, 21\}$. The values at the nonzero coordinates are obtained from normalizing (with respect to Σ) uniform random integers drawn from $\{-2, 1, 0, 1, 2\}$. The details of the four settings are as follows:

1. `Identity`: We set $\Sigma = I_p$. Since the PMA approach implicitly assumes that both Σ_x and Σ_y are identity matrices, this setting is to its favor.

2. `Toeplitz`: We set

$$\Sigma_{ij} = 0.3^{|i-j|}, \quad i, j \in [p].$$

   In other words, Σ_x and Σ_y are Toeplitz matrices.

3. `SparseInv`: We let $\Sigma = \Omega^{-1}$ with

$$\Omega_{ij} = \mathbf{1}_{\{i=j\}} + 0.5 \times \mathbf{1}_{\{|i-j|=1\}} + 0.4 \times \mathbf{1}_{\{|i-j|=2\}}, \quad i, j \in [p].$$

   In other words, Σ_x and Σ_y have sparse inverse matrices.

4. `Dense`: We let $\Sigma = (\sigma^0_{ij}/\sqrt{\sigma^0_{ii}\sigma^0_{jj}})$ where $\Sigma^0 = (\sigma^0_{ij}) = I_p + W_p/20$ with W_p a random matrix generated from the Wishart distribution $W_p(20, I_p)$.

Results Tables 1 – 4 report, in each of the four settings, the medians and the median absolute deviations (MADs) of the estimation errors of the proposed method and of the PMA method out of 100 repetitions for three different configurations of (p, m, n) values. From the simulation results, our method consistently outperform the PMA method by a large margin. It is worth noting that even in the `Identity` setting, which should favor the PMA approach, our method still leads to much smaller estimation errors. In the other three settings, the advantage of our method is more substantial. Comparing the first and the second blocks in Tables 1 – 4, we see that for the same settings, larger sample size leads to more accurate estimation. Comparing the second and the third blocks in Tables 1 – 4, we see that for the same sparsity levels and

(p, m, n)	Method	$\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}$	$\|P_{\widehat{V}} - P_V\|_{\mathrm{F}}$
(200, 200, 500)	CoLaR	0.150 (0.025)	0.160 (0.029)
	PMA	0.444 (0.054)	0.428 (0.055)
(200, 200, 750)	CoLaR	0.110 (0.018)	0.120 (0.018)
	PMA	0.353 (0.031)	0.334 (0.045)
(500, 500, 750)	CoLaR	0.126 (0.021)	0.138 (0.021)
	PMA	0.407 (0.187)	0.445 (0.246)

Table 1: Estimation errors (`Identity`): Median and MAD (in parentheses) in 100 repetitions.

the same sample sizes, the estimation errors are not too sensitive with respect to the ambient dimension, which is consistent with the theoretical results in Section 3. Last but not least, comparing the four tables, we find that the proposed method does not seem to be too sensitive to the underlying covariance structure Σ_x and Σ_y. In summary, the proposed method delivers consistent and competitive performance in all the three covariance settings across all dimension and sample size configurations, and its behavior agrees well with the theoretical results.

5 Real Data Example

To further demonstrate the potential application of the proposed method, we present its result on a breast cancer dataset in [21]. The dataset records both the DNA methylation and gene expression data for 99 breast cancer patients that belong to the "Luminal A" subtype as determined in [21].

We first apply the same screening approach as in [9] to select 74 genes and 1600 methylation probes distributed on 22 chromosomes. To be specific, we applied a marginal logistic regression with the disease-free status variable for each gene and each

(p, m, n)	Method	$\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}$	$\|P_{\widehat{V}} - P_V\|_{\mathrm{F}}$
(200, 200, 500)	CoLaR	0.146 (0.020)	0.159 (0.025)
	PMA	0.627 (0.076)	0.581 (0.070)
(200, 200, 750)	CoLaR	0.113 (0.021)	0.123 (0.016)
	PMA	0.571 (0.042)	0.561 (0.068)
(500, 500, 750)	CoLaR	0.133 (0.020)	0.139 (0.023)
	PMA	0.597 (0.139)	0.586 (0.182)

Table 2: Estimation errors (`Toeplitz`): Median and MAD (in parentheses) in 100 repetitions.

(p, m, n)	Method	$\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}$	$\|P_{\widehat{V}} - P_V\|_{\mathrm{F}}$
(200, 200, 500)	CoLaR	0.143 (0.019)	0.187 (0.033)
	PMA	1.560 (0.046)	1.685 (0.072)
(200, 200, 750)	CoLaR	0.106 (0.019)	0.143 (0.028)
	PMA	1.567 (0.026)	1.701 (0.047)
(500, 500, 750)	CoLaR	0.110 (0.016)	0.167 (0.035)
	PMA	1.705 (0.023)	1.710 (0.061)

Table 3: Estimation errors (`SparseInv`): Median and MAD (in parentheses) in 100 repetitions.

(p, m, n)	Method	$\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}$	$\|P_{\widehat{V}} - P_V\|_{\mathrm{F}}$
(200, 200, 500)	CoLaR	0.171 (0.028)	0.198 (0.030)
	PMA	1.031 (0.041)	0.894 (0.039)
(200, 200, 750)	CoLaR	0.135 (0.019)	0.152 (0.022)
	PMA	1.001 (0.029)	0.882 (0.048)
(500, 500, 750)	CoLaR	0.135 (0.020)	0.164 (0.022)
	PMA	1.025 (0.027)	0.803 (0.027)

Table 4: Estimation errors (`Dense`): Median and MAD (in parentheses) in 100 repetitions.

methylation, respectively. The selected 74 genes and 1600 methylation probes have p-values less than 0.01. To further control the ambient dimensions of the datasets, we apply CoLaR to 74 genes and the methylation probes on each chromosome separately. To remove false discovery, for each chromosome, we randomly select 66 out of the 99 patients as training set and the remaining 33 patients as test set. We apply CoLaR on the training set to obtain estimates of U and V, and then project the test set on the estimated canonical correlation directions to compute the canonical correlation on the test set. Fig. 1 includes the boxplots of canonical correlations on test datasets based on 25 random splits of the training and test datasets, where we applied CoLaR with $r = 1$ and $0.5\sqrt{\log(p \vee m)/n}$ and $0.5\sqrt{(r + \log(p \vee m))/n}$ as penalty parameters in the first and the second stages of the method.

From the boxplots, Chromosomes 2, 4, 10 and 19 have all 25 test data canonical correlations greater than 0.2. Thus, we applied CoLaR with the foregoing specified parameters to all the 99 samples on these four chromosomes. In Table 5, we report for each of the four chromosomes the number of methylation probes after screening and the five genes and methylation probes which have the largest absolute values in the estimated canonical correlation direction vectors. We notice among the four

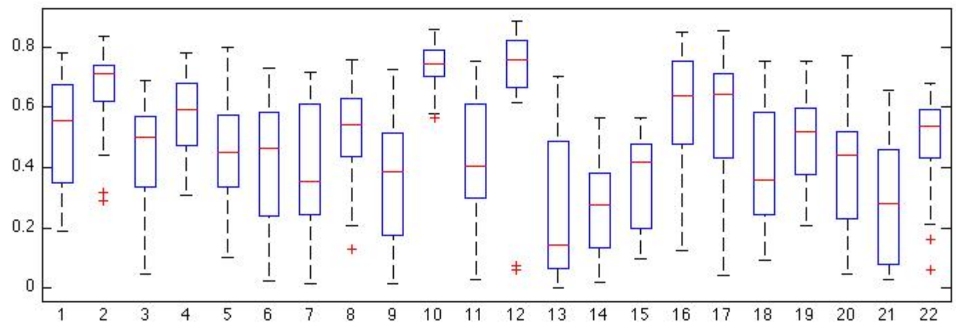

Figure 1: Boxplots of canonical correlations on test datasets based on 25 runs.

genes, MBNL1, CCL15, MEOX2 and EMCN, at least three of them appear in all four chromosomes. These genes are reported and studied by [5, 2, 17, 18] in the literature of breast cancer research.

6 Proofs

In this section, we present proofs of the theorems in Section 3. Note that the proofs of Theorem 3.1 and Theorem 3.2 are essentially independent. Thus, the same symbol used in the proofs of Theorems 3.1 and 3.2 can represent different quantities. Proofs of the technical lemmas used in this section are deferred to Section 7.

6.1 Proof of Theorem 3.1

Before stating the proof, let us introduce some notation and technical lemmas. Define

$$\begin{aligned} \widetilde{U} &= U(U'\widehat{\Sigma}_x U)^{-1/2}, \quad \widetilde{V} = V(V'\widehat{\Sigma}_y V)^{-1/2}, \\ \widetilde{A} = \widetilde{U}\widetilde{V}', \quad \widetilde{\Lambda} &= (U'\widehat{\Sigma}_x U)^{1/2}\Lambda(V'\widehat{\Sigma}_y V)^{1/2}, \quad A = UV'. \end{aligned} \tag{24}$$

The reason for defining these quantities is because $\widehat{\Sigma}_x^{1/2}\widetilde{U} \in O(p,r)$ and $\widehat{\Sigma}_y^{1/2}\widetilde{V} \in O(m,r)$, which facilitates the proof. Due to the sparsity of U and V, the matrices

Chromosome	# probes	Top genes / methylation probes
2	92	MBNL1, CCL15, MEOX2, EMCN, REEP2 cg02251243, cg07683388, cg09694782, cg26132737, cg22115977
4	53	PRKCH, MBNL1, MEOX2, CCL15, IL33 cg15919816, cg06663149, cg25986240, cg06059810, cg06767059
10	62	MEOX2, EMCN, THSD7A, IL33, CCL15 cg02859866, cg01088382, cg11612727, cg12627983, cg13846998
19	113	EMCN, MEOX2, IL33, MBNL1, NR0B1 cg00431565, cg05562817, cg24731702, cg19577671, cg27659109

Table 5: Top genes and methylation probes on Chromosomes 2, 4, 10 and 19.

$\widetilde{U}, \widetilde{V}, \widetilde{A}, \widetilde{\Lambda}$ are good approximations to U, V, A, Λ. This is established rigorously in the following lemma.

Lemma 6.1. *Assume* $\frac{1}{n}(s_u + s_v + \log(ep/s_u) + \log(em/s_v)) \leq C_1$ *for some constant* $c > 0$*. Then, for any* $C' > 0$*, there exists* $C > 0$ *only depending on* C' *such that*

$$\begin{aligned}
\|\widetilde{U} - U\|_{\mathrm{op}} &\leq C\sqrt{\frac{1}{n}\left(s_u + \log\frac{ep}{s_u}\right)}, \\
\|\widetilde{V} - V\|_{\mathrm{op}} &\leq C\sqrt{\frac{1}{n}\left(s_v + \log\frac{em}{s_v}\right)}, \\
\|\widetilde{A} - A\|_{\mathrm{op}} \vee \|\widetilde{\Lambda} - \Lambda\|_{\mathrm{op}} &\leq C\left[\sqrt{\frac{1}{n}\left(s_u + \log\frac{ep}{s_u}\right)} + \sqrt{\frac{1}{n}\left(s_v + \log\frac{em}{s_v}\right)}\right],
\end{aligned}$$

with probability at least $1 - \exp(-C'(s_u + \log(ep/s_u))) - \exp(C'(s_v + \log(em/s_v)))$*.*

Note that Lemma 6.1 also implies the existence of $\widetilde{U}, \widetilde{V}, \widetilde{A}, \widetilde{\Lambda}$ by ensuring that $U'\widehat{\Sigma}_y U$ and $V'\widehat{\Sigma}_y V$ are invertible with high probability (see Lemma 7.1). The next lemma shows the matrix $\widetilde{A}$, which serves as a surrogate of the truth A, is in the feasible set of the program (9).

Lemma 6.2. *When $\widetilde{A}$ exists, we have*

$$\|\widehat{\Sigma}_x^{1/2}\widetilde{A}\widehat{\Sigma}_y^{1/2}\|_* = r \quad \text{and} \quad \|\widehat{\Sigma}_x^{1/2}\widetilde{A}\widehat{\Sigma}_y^{1/2}\|_{\mathrm{op}} = 1.$$

The following lemma characterizes the curvature of the objective function. It is comparable to Lemma 9 in [11]. The difference is that we allow a non-diagonal K and a more general E.

Lemma 6.3. *Let $F \in O(p,r)$, $G \in O(m,r)$ and $K \in \mathbb{R}^{r\times r}$ with positive diagonal elements $\{k_{ll}\}_{l=1}^r$. If E satisfies $\|E\|_{\mathrm{op}} \le 1$ and $\|E\|_* \le r$, then*

$$\langle FKG', FG' - E\rangle \ge \frac{\min_{1\le l\le r} k_{ll}}{2}\|FG' - E\|_{\mathrm{F}}^2. \tag{25}$$

The requirement on E in Lemma 6.3 are that $\|E\|_{\mathrm{op}} \le 1$ and that $\|E\|_* \le r$, which coincide with the two constraints in the program (9), respectively. Next, define

$$\widetilde{\Sigma}_{xy} = \widehat{\Sigma}_x U \Lambda V' \widehat{\Sigma}_y. \tag{26}$$

The following lemma shows $\widetilde{\Sigma}_{xy}$ is close to $\widehat{\Sigma}_{xy}$ uniformly over each entry.

Lemma 6.4. *Assume $r\sqrt{\frac{\log(p+m)}{n}} \le C_1$ for some constant $C_1 > 0$. Then, for any $C' > 0$, there exists a constant $C > 0$ only depending on C_1, C', M, such that*

$$||\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}||_\infty \le C\sqrt{\frac{\log(p+m)}{n}},$$

with probability at least $1-(p+m)^{-C'}$.

Note that the assumption $r\sqrt{\frac{\log(p+m)}{n}} \le C_1$ is always implied by (22) because $r \le s_u \wedge s_v$. Finally, we need a lemma on restricted eigenvalue. For any p.s.d. matrix B, define

$$\phi^B_{\max}(k) = \max_{||u||_0\le k, u\neq 0}\frac{u'Bu}{u'u}, \quad \phi^B_{\min}(k) = \min_{||u||_0\le k, u\neq 0}\frac{u'Bu}{u'u}.$$

The following lemma is adapted from Lemma 14 in [11]. The original Lemma 14 in [11] is stated for the Gaussian case. The result also applies to the sub-Gaussian case with the same proof.

Lemma 6.5. *Assume* $\frac{1}{n}\big((k_u \wedge p)\log(ep/(k_u \wedge p)) + (k_v \wedge m)\log(em/(k_v \wedge m))\big) \leq C_1$ *for some constant* $C_1 > 0$*. Then, for any* $C' > 0$*, there exists a constant* $C > 0$ *only depending on* C_1, C', M*, such that for* $\delta_u(k_u) = \sqrt{\frac{(k_u \wedge p)\log(ep/(k_u \wedge p))}{n}}$ *and* $\delta_v(k_v) = \sqrt{\frac{(k_v \wedge m)\log(em/(k_v \wedge m))}{n}}$*, we have*

$$M^{-1} - C\delta_u(k_u) \leq \phi_{\min}^{\widehat{\Sigma}_x}(k_u) \leq \phi_{\max}^{\widehat{\Sigma}_x}(k_u) \leq M + C\delta_u(k_u),$$
$$M^{-1} - C\delta_v(k_v) \leq \phi_{\min}^{\widehat{\Sigma}_y}(k_v) \leq \phi_{\max}^{\widehat{\Sigma}_y}(k_v) \leq M + C\delta_v(k_v),$$

with probability at least $1 - \exp\big(-C'(k_u \wedge p)\log(ep/(k_u \wedge p))\big) - \exp\big(-C'(k_v \wedge m)\log(em/(k_v \wedge m))\big)$.

Now we are ready to state the proof of Theorem 3.1.

Proof of Theorem 3.1. The proof consists of three steps. In the first step, we are going to derive a bound for $\|\widehat{\Sigma}_x^{1/2}(\widehat{A}-\widetilde{A})\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$. In the second step, we derive a cone condition and use it to lower bound $\|\widehat{\Sigma}_x^{1/2}(\widehat{A}-\widetilde{A})\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$ by a constant multiple of $\|\widehat{A}-\widetilde{A}\|_{\mathrm{F}}$. Finally, in the third step, we use Wedin's sin-theta theorem [29] to show that the bound for $\|\widehat{A}-\widetilde{A}\|_{\mathrm{F}}$ implies a bound for $\|P_{U^{(0)}} - P_U\|_{\mathrm{F}} \vee \|P_{V^{(0)}} - P_V\|_{\mathrm{F}}$.

Step 1. Recall $\widetilde{A}$ in (24). By Lemma 6.1, $\widetilde{A}$ is well-defined with high probability and feasible with respect to the program (9) according to Lemma 6.2. Then, by the definition of $\widehat{A}$, we have

$$\left\langle \widehat{\Sigma}_{xy}, \widehat{A} \right\rangle - \rho||\widehat{A}||_1 \geq \left\langle \widehat{\Sigma}_{xy}, \widetilde{A} \right\rangle - \rho||\widetilde{A}||_1.$$

After rearrangement, we have

$$-\left\langle \widetilde{\Sigma}_{xy}, \Delta \right\rangle \leq \rho\Big(||\widetilde{A}||_1 - ||\widetilde{A}+\Delta||_1\Big) + \left\langle \widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}, \Delta \right\rangle, \tag{27}$$

where $\widetilde{\Sigma}_{xy}$ is defined in (26), and $\Delta = \widehat{A} - \widetilde{A}$. For the first term on the right hand side of (27), we have

$$\begin{aligned} ||\widetilde{A}||_1 - ||\widetilde{A}+\Delta||_1 &= ||\widetilde{A}_{S_uS_v}||_1 - ||\widetilde{A}_{S_uS_v} + \Delta_{S_uS_v}||_1 - ||\Delta_{(S_uS_v)^c}||_1 \\ &\leq ||\Delta_{S_uS_v}||_1 - ||\Delta_{(S_uS_v)^c}||_1. \end{aligned}$$

For the second term on the right hand side of (27), we have $\left\langle \widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}, \Delta \right\rangle \le ||\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}||_\infty ||\Delta||_1$. Thus when

$$\rho \ge 2||\widehat{\Sigma}_{xy} - \widehat{\Sigma}_{xy}||_\infty, \tag{28}$$

we have

$$-\left\langle \widetilde{\Sigma}_{xy}, \Delta \right\rangle \le \frac{3\rho}{2}||\Delta_{S_u S_v}||_1 - \frac{\rho}{2}||\Delta_{(S_u S_v)^c}||_1. \tag{29}$$

Using Lemma 6.3 and the definition (24), we can lower bound the left hand side of (29) as

$$\begin{aligned} -\left\langle \widetilde{\Sigma}_{xy}, \Delta \right\rangle &= \left\langle \widehat{\Sigma}_x^{1/2} U \Lambda V' \widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2} (\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2} \right\rangle \\ &= \left\langle \widehat{\Sigma}_x^{1/2} \widetilde{U} \widetilde{\Lambda} \widetilde{V}' \widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2} (\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2} \right\rangle \\ &\ge \frac{1}{2} \min_{1 \le l \le r} \widetilde{\lambda}_{ll} \|\widehat{\Sigma}_x^{1/2} (\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2, \end{aligned}$$

where $\widetilde{\lambda}_{ll}$ is the (l, l)-th entry of $\widetilde{\Lambda}$. Using Lemma 6.1 and the assumption (22), we have

$$\min_{1 \le l \le r} \widetilde{\lambda}_{ll} \ge \lambda_r - ||\widetilde{\Lambda} - \Lambda||_\infty \ge \lambda_r - \|\widetilde{\Lambda} - \Lambda\|_{\mathrm{op}} \ge \frac{1}{2}\lambda_r,$$

with high probability. Hence, we have

$$-\left\langle \widetilde{\Sigma}_{xy}, \Delta \right\rangle \ge \frac{1}{4}\lambda_r \|\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2. \tag{30}$$

Moreover, the right hand side of (29) can be upper bounded by

$$\frac{3\rho}{2}||\Delta_{S_u S_v}||_1 \le \frac{3\sqrt{s_u s_v}}{2}\rho \|\Delta_{S_u S_v}\|_{\mathrm{F}}.$$

Combining this with (30), we have

$$\lambda_r \|\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 \le 6\sqrt{s_u s_v}\rho \|\Delta_{S_u S_v}\|_{\mathrm{F}}, \tag{31}$$

which completes the first step.

Step 2. Combining (29) and (30), we obtain the cone condition

$$||\Delta_{(S_u S_v)^c}||_1 \le 3||\Delta_{S_u S_v}||_1. \tag{32}$$

Motivated by the argument in [6], let the index set $J_1 = \{(i_k, j_k)\}_{k=1}^t$ in $(S_u \times S_v)^c$ correspond to the entries with the largest absolute values in Δ, and we define the set $\widetilde{J} = (S_u \times S_v) \cup J_1$. Now we partition $\widetilde{J}^c$ into disjoint subsets $J_2, ..., J_K$ of size t (with $|J_K| \leq t$), such that J_k is the set of (double) indices corresponding to the entries of t largest absolute values in Δ outside $\widetilde{J} \cup \bigcup_{j=2}^{k-1} J_j$. By triangle inequality,

$$
\begin{aligned}
&\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} \\
&\geq \|\widehat{\Sigma}_x^{1/2}\Delta_{\widetilde{J}}\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} - \sum_{k=2}^{K}\|\widehat{\Sigma}_x^{1/2}\Delta_{J_k}\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} \\
&\geq \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u+t)\phi_{\min}^{\widehat{\Sigma}_y}(s_v+t)}\|\Delta_{\widetilde{S}_u\widetilde{S}_v}\|_{\mathrm{F}} - \sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)\phi_{\max}^{\widehat{\Sigma}_y}(t)}\sum_{k=2}^{K}\|\Delta_{J_k}\|_{\mathrm{F}}.
\end{aligned}
$$

By the construction of J_k, we have

$$
\begin{aligned}
&\sum_{k=2}^{K}\|\Delta_{J_k}\|_{\mathrm{F}} \\
&\leq \sqrt{t}\sum_{k=2}^{K}||\Delta_{J_k}||_\infty \leq t^{-1/2}\sum_{k=2}^{K}||\Delta_{J_{k-1}}||_1 \leq t^{-1/2}||\Delta_{(S_uS_v)^c}||_1 \\
&\leq 3t^{-1/2}||\Delta_{S_uS_v}||_1 \leq 3\sqrt{\frac{s_us_v}{t}}\|\Delta_{S_uS_v}\|_{\mathrm{F}} \leq 3\sqrt{\frac{s_us_v}{t}}\|\Delta_{\widetilde{J}}\|_{\mathrm{F}}, \qquad (33)
\end{aligned}
$$

where we have used the cone condition (32). Hence, we have the lower bound

$$
\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} \geq \kappa\|\Delta_{\widetilde{J}}\|_{\mathrm{F}},
$$

with

$$
\kappa = \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u+t)\phi_{\min}^{\widehat{\Sigma}_y}(s_v+t)} - 3\sqrt{\frac{s_us_v}{t}}\sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)\phi_{\max}^{\widehat{\Sigma}_y}(t)}. \qquad (34)
$$

Taking $t = c_1 s_u s_v$ for some sufficiently large constant $c_1 > 1$, with high probability, κ can be lower bounded by a positive constant κ_0 only depending on M. To see this, note that by Lemma 6.5, (34) can be lower bounded by the difference of $\sqrt{M^{-1} - C\delta_u(2c_1s_us_v)}\sqrt{M^{-1} - C\delta_v(2c_1s_us_v)}$ and $3c_1^{-1/2}\sqrt{M + C\delta_u(c_1s_us_v)}\sqrt{M + C\delta_v(c_1s_us_v)}$ where δ_u and δ_v are defined as in Lemma 6.5. It is sufficient to show that $\delta_u(2c_1s_us_v)$,

$\delta_v(2c_1s_us_v)$, $\delta_u(c_1s_us_v)$ and $\delta_v(c_1s_us_v)$ are sufficiently small to get a positive absolute constant κ_0. For the first term, when $2c_1s_us_v \le p$, it is bounded by $\frac{2c_1s_us_v\log(ep)}{n}$ and is sufficiently small under the assumption (22). When $2c_1s_us_v > p$, it is bounded by $\frac{2c_1s_us_v}{n}$ and is also sufficiently small under (22). The same argument also holds for the other terms.

Together with (31), this brings the bound

$$\|\Delta_{\widetilde{J}}\|_{\mathrm{F}} \le \frac{C\sqrt{s_us_v}\rho}{\kappa_0^2\lambda_r}. \tag{35}$$

By (33), we have

$$\|\Delta_{\widetilde{J}^c}\|_{\mathrm{F}} \le \sum_{k=2}^{K}\|\Delta_{J_k}\|_{\mathrm{F}} \le 3\sqrt{\frac{s_us_v}{t}}\|\Delta_{\widetilde{J}}\|_{\mathrm{F}} \le 3c_1^{-1/2}\|\Delta_{\widetilde{J}}\|_{\mathrm{F}}. \tag{36}$$

Summing (35) and (36), we have $\|\Delta\|_{\mathrm{F}} \le C\frac{\sqrt{s_us_v}\rho}{\lambda_r}$ with high probability. According to Lemma 6.4, we may choose $\rho \ge \gamma\sqrt{\frac{\log(p+m)}{n}}$ so that (28) holds with high probability. Hence,

$$\|\Delta\|_{\mathrm{F}} \le C\frac{\sqrt{s_us_v}\rho}{\lambda_r}, \tag{37}$$

with high probability. This completes the second step.

Step 3. By Wedin's sin-theta theorem [29], we have

$$\|P_{U^{(0)}} - P_U\|_{\mathrm{F}} = \|P_{U^{(0)}} - P_{\widetilde{U}}\|_{\mathrm{F}} \le \frac{C\|\widehat{A}-\widetilde{A}\|_{\mathrm{F}}}{\sigma_r(\widehat{A})-\sigma_{r+1}(\widetilde{A})},$$

where $\sigma_{r+1}(\widetilde{A}) = 0$ because $\widetilde{A}$ is a rank-r matrix. Using Weyl's inequality [12, p.449], we lower bound $\sigma_r(\widehat{A})$ by

$$\begin{aligned}\sigma_r(\widehat{A}) &\ge \sigma_r(UV') - \|\widehat{A}-\widetilde{A}\|_{\mathrm{op}} - \|\widetilde{U}\widetilde{V}' - UV'\|_{\mathrm{op}} \\ &\ge \sigma_r(UV') - \|\widehat{A}-\widetilde{A}\|_{\mathrm{F}} - \|\widetilde{U}\widetilde{V}' - UV'\|_{\mathrm{op}}.\end{aligned}$$

Since $\Sigma_x^{1/2}U \in O(p,r)$ and $\Sigma_y^{1/2}V \in O(m,r)$, $\sigma_r(UV')$ is at a constant level. By (37) and Lemma 6.1, $\|\widehat{A}-\widetilde{A}\|_{\mathrm{F}}$ and $\|\widetilde{U}\widetilde{V}' - UV'\|_{\mathrm{op}}$ are sufficiently small with high probability.

Hence, $\sigma_r(\widehat{A})$ is bounded below by a constant and

$$\|P_{U^{(0)}} - P_U\|_{\mathrm{F}} \leq C\frac{\sqrt{s_u s_v}\rho}{\lambda_r}.$$

The same bound holds for $\|P_{V^{(0)}} - P_V\|_{\mathrm{F}}$ by a similar argument. Finally, $\|\widehat{A} - A\|_{\mathrm{F}}$ can be bounded by the simple inequality

$$\|\widehat{A} - A\|_{\mathrm{F}} \leq \|\widehat{A} - \widetilde{A}\|_{\mathrm{F}} + \sqrt{2r}\|\widetilde{U}\widetilde{V}' - UV'\|_{\mathrm{op}},$$

where the first term is bounded by (37), and the second term is bounded by the desired rate using Lemma 6.1 and the fact $r \leq s_u \wedge s_v$. Hence, $\|\widehat{A} - A\|_{\mathrm{F}} \leq C\frac{\sqrt{s_u s_v}\rho}{\lambda_r}$. The proof is complete by applying a union bound to all probabilistic argument we have made. □

6.2 Proof of Theorem 3.2

Define

$$U^* = U\Lambda V'\Sigma_y V^{(0)}, \quad \Delta = \widehat{U} - U^*. \tag{38}$$

Note that Δ is different from the one used in the proof of Theorem 3.1.

Lemma 6.6. *Assume $\frac{r+\log p}{n} \leq C_1$ for some constant $C_1 > 0$. Then, for any $C' > 0$, there exists a constant $C > 0$ only depending on C_1, C', M, such that*

$$\max_{1\leq j\leq p} ||[\widehat{\Sigma}_{xy}V^{(0)} - \widehat{\Sigma}_x U^*]_{j\cdot}|| \leq C\sqrt{\frac{r + \log p}{n}},$$

with probability at least $1 - \exp\big(-C'(r + \log p)\big)$.

Proof of Theorem 3.2. Since the analysis for $\widehat{U}$ and $\widehat{V}$ are the same, we only state the proof for $\widehat{U}$. The proof consists of three steps. In the first step, we derive a bound for $\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta)$. In the second step, we derive a cone condition and use it to obtain a bound for $\|\Delta\|_{\mathrm{F}}$ by arguing that $\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta)$ upper bounds $\|\Delta\|_{\mathrm{F}}$. In the third step, a sin-theta theorem is applied to bound $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}$ by $\|\Delta\|_{\mathrm{F}}$.

Step 1. By definition of $\widehat{U}$, we have

$$\begin{aligned} & \mathsf{Tr}(\widehat{U}'\widehat{\Sigma}_x\widehat{U}) - 2\,\mathsf{Tr}(\widehat{U}'\widehat{\Sigma}_{xy}V^{(0)}) + \rho_u \sum_{j=1}^{p} ||\widehat{U}_{j\cdot}|| \\ \leq\ & \mathsf{Tr}((U^*)'\widehat{\Sigma}_x U^*) - 2\,\mathsf{Tr}((U^*)'\widehat{\Sigma}_{xy}V^{(0)}) + \rho_u \sum_{j=1}^{p} ||U^*_{j\cdot}||. \end{aligned}$$

After rearrangement, we have

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \leq \rho_u \sum_{j=1}^{p} \Big(||U^*_{j\cdot}|| - ||U^*_{j\cdot} + \Delta_{j\cdot}|| \Big) + 2\,\mathsf{Tr}\Big(\Delta'(\widehat{\Sigma}_{xy}V^{(0)} - \widehat{\Sigma}_x U^*)\Big). \tag{39}$$

For the first term on the right hand side of (39), we have

$$\begin{aligned} & \sum_{j=1}^{p} \Big(||U^*_{j\cdot}|| - ||U^*_{j\cdot} + \Delta_{j\cdot}|| \Big) \\ =\ & \sum_{j\in S_u} ||U^*_{j\cdot}|| - \sum_{j\in S_u} ||U^*_{j\cdot} - \Delta_{j\cdot}|| - \sum_{j\in S_u^c} ||\Delta_{j\cdot}|| \\ \leq\ & \sum_{j\in S_u} ||\Delta_{j\cdot}|| - \sum_{j\in S_u^c} ||\Delta_{j\cdot}||. \end{aligned}$$

For the second term on the right hand side of (39), we have

$$\begin{aligned} & \mathsf{Tr}\Big(\Delta'(\widehat{\Sigma}_{xy}V^{(0)} - \widehat{\Sigma}_x U^*)\Big) \\ \leq\ & \Big(\sum_{j=1}^{p} ||\Delta_{j\cdot}|| \Big) \max_{1\leq j\leq p} ||[\widehat{\Sigma}_{xy}V^{(0)} - \widehat{\Sigma}_x U^*]_{j\cdot}||, \end{aligned}$$

where $[\cdot]_{j\cdot}$ means the j-th row of the corresponding matrix. When

$$\rho_u \geq 4 \max_{1\leq j\leq p} ||[\widehat{\Sigma}_{xy}V^{(0)} - \widehat{\Sigma}_x U^*]_{j\cdot}||, \tag{40}$$

we have

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \leq \frac{3\rho_u}{2} \sum_{j\in S_u} ||\Delta_{j\cdot}|| - \frac{\rho_u}{2} \sum_{j\in S_u^c} ||\Delta_{j\cdot}||. \tag{41}$$

Since $\sum_{j\in S_u} ||\Delta_{j\cdot}|| \leq \sqrt{s_u}\sqrt{\sum_{j\in S_u} ||\Delta_{j\cdot}||^2}$, (41) can be upper bounded by

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \leq \frac{3\sqrt{s_u}\rho_u}{2} \sqrt{\sum_{j\in S_u} ||\Delta_{j\cdot}||^2}. \tag{42}$$

This completes the first step.

Step 2. The inequality (41) implies the cone condition

$$\sum_{j\in S_u^c}||\Delta_{j\cdot}||\leq 3\sum_{j\in S_u}||\Delta_{j\cdot}||. \tag{43}$$

Let the index set $J_1=\{j_1,...,j_t\}$ in S_u^c correspond to the rows with the largest l_2 norm in Δ, and we define the extended support $\widetilde{S}_u=S_u\cup J_1$. Now we partition $\widetilde{S}_u^c$ into disjoint subsets $J_2,...,J_K$ of size t (with $|J_K|\leq t$), such that J_k is the set of indices corresponding to the t rows with largest l_2 norm in Δ outside $\widetilde{S}_u\cup\bigcup_{j=2}^{k-1}J_j$. Note that $\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta)=\|n^{-1/2}X\Delta\|_{\rm F}^2$, where $X=[X_1,...,X_n]'\in\mathbb{R}^{n\times p}$ denotes the data matrix. By triangle inequality, we have

$$\begin{aligned}\|n^{-1/2}X\Delta\|_{\rm F} &\geq \|n^{-1/2}X\Delta_{\widetilde{S}_u*}\|_{\rm F}-\sum_{k\geq 2}\|n^{-1/2}X\Delta_{J_k*}\|_{\rm F}\\ &\geq \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u+t)}\|\Delta_{\widetilde{S}_u*}\|_{\rm F}-\sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)}\sum_{k\geq 2}\|\Delta_{J_k*}\|_{\rm F},\end{aligned}$$

where for a subset $B\subset[p]$, $\Delta_{B*}=(\Delta_{ij}\mathbf{1}_{\{i\in B,j\in[r]\}})$, and

$$\begin{aligned}\sum_{k\geq 2}\|\Delta_{J_k*}\|_{\rm F} &\leq \sqrt{t}\sum_{k\geq 2}\max_{j\in J_k}||\Delta_{j\cdot}||\leq\sqrt{t}\sum_{k\geq 2}\frac{1}{t}\sum_{j\in J_{k-1}}||\Delta_{j\cdot}|| \qquad (44)\\ &\leq t^{-1/2}\sum_{j\in S_u^c}||\Delta_{j\cdot}||\leq 3t^{-1/2}\sum_{j\in S_u}||\Delta_{j\cdot}||\\ &\leq 3\sqrt{\frac{s_u}{t}}\sqrt{\sum_{j\in S_u}||\Delta_{j\cdot}||^2}\leq 3\sqrt{\frac{s_u}{t}}\|\Delta_{\widetilde{S}_u*}\|_{\rm F}. \qquad (45)\end{aligned}$$

In the above derivation, we have used the construction of J_k and the cone condition (43). Hence,

$$\|n^{-1/2}X\Delta\|_{\rm F}\geq\kappa\|\Delta_{\widetilde{S}_u*}\|_{\rm F},$$

with $\kappa=\sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u+t)}-3\sqrt{\frac{s_u}{t}}\sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)}$. In view of Lemma 6.5, taking $t=c_1s_u$ for some sufficiently large constant c_1, with high probability, κ can be lower bounded by a positive constant κ_0 only depending on M. Combining with (42), we have

$$\|\Delta_{\widetilde{S}_u*}\|_{\rm F}\leq\frac{C\sqrt{s_u}\rho_u}{2\kappa_0^2}. \tag{46}$$

By (44)-(45), we have

$$\|\Delta_{(\widetilde{S}_u)^c*}\|_{\rm F} \leq \sum_{k\geq 2} \|\Delta_{J_k*}\|_{\rm F} \leq 3\sqrt{\frac{s_u}{t}}\|\Delta_{\widetilde{S}_u*}\|_{\rm F} \leq 3c_1^{-1/2}\|\Delta_{\widetilde{S}_u*}\|_{\rm F}. \tag{47}$$

Summing (46) and (47), we have $\|\Delta\|_{\rm F} \leq C\sqrt{s_u}\rho$. According to Lemma 6.6, we may choose $\rho_u \geq \gamma_u\sqrt{\frac{r+\log p}{n}}$ so that (40) holds with high probability. Hence,

$$\|\Delta\|_{\rm F} \leq C\sqrt{\frac{s_u(r+\log p)}{n}}, \tag{48}$$

with high probability. This completes the second step.

Step 3. By Wedin's sin-theta theorem [29], we have

$$\|P_{\widehat{U}} - P_U\|_{\rm F} = \|P_{\widehat{U}} - P_{U^*}\|_{\rm F} \leq \frac{C\|\widehat{U} - U^*\|_{\rm F}}{\sigma_r(U^*) - \sigma_{r+1}(\widehat{U})}.$$

Since $\widehat{U} \in \mathbb{R}^{p\times r}$, $\sigma_{r+1}(\widehat{U}) = 0$. We lower bound $\sigma_r(U^*)$ by

$$\sigma_r(U^*) \geq C^{-1}\lambda_r\sigma_{\min}(V'\Sigma_y V^{(0)}).$$

Since $\|\widehat{U} - U^*\|_{\rm F}$ is upper bounded by (48), we have

$$\|P_{\widehat{U}} - P_U\|_{\rm F} \leq C\frac{\sqrt{s_u}\rho_u}{\lambda_r\sigma_{\min}(V'\Sigma_y V^{(0)})},$$

with high probability. A similar argument gives

$$\|P_{\widehat{V}} - P_V\|_{\rm F} \leq C\frac{\sqrt{s_v}\rho_v}{\lambda_r\sigma_{\min}(U'\Sigma_x U^{(0)})}.$$

Hence, the proof is complete. □

6.3 Proof of Theorem 3.3

To facilitate the proof, we need the following result.

Lemma 6.7 (Stewart and Sun [23], Theorem II.4.11)**.** *For any matrices F, G with $F'F = G'G = I_r$, we have*

$$\inf_{W\in O(r,r)} \|F - GW\|_{\rm F} \leq \|FF' - GG'\|_{\rm F}.$$

Proof of Theorem 3.3. It is sufficient to lower bound $\sigma_{\min}(V'\Sigma_y V^{(0)})$ and $\sigma_{\min}(U'\Sigma_x U^{(0)})$ by constants. Let V have singular value decomposition $V = RDQ'$. By Lemma 6.7 and Theorem 3.1, there exists a matrix $W \in O(r, r)$, such that

$$\|V^{(0)} - RW\|_{\text{op}} \le \|V^{(0)} - RW\|_{\text{F}} \le C\sqrt{\frac{s_u s_v \log(p+m)}{n\lambda_r^2}}, \tag{49}$$

with high probability. By Weyl's inequality,

$$\sigma_{\min}(V'\Sigma_y V^{(0)}) \ge \sigma_{\min}(V'\Sigma_y RW) - \|V'\Sigma_y(V^{(0)} - RW)\|_{\text{op}}. \tag{50}$$

Combining (49), (50) and the assumption (22), it is sufficient to lower bound $\sigma_{\min}(V'\Sigma_y RW)$ by a constant. Note that $V'\Sigma_y RW = V'\Sigma_y VQD^{-1}W = QD^{-1}W$, and thus we have

$$\sigma_{\min}(V'\Sigma_y RW) = \sigma_{\min}(QD^{-1}W) = \sigma_{\min}(D^{-1}) = \|V\|_{\text{op}}^{-1} \ge M^{-1/2}.$$

Applying the same argument for $\sigma_{\min}(U'\Sigma_x U^{(0)})$, the proof is complete. □

6.4 Proof of Theorem 3.4

The proof largely follows the proof of Theorem 3 in [11], though [11] considered a different loss function from the current paper. Nonetheless, we spell out the details below for the sake of completeness.

For any probability measures $\mathbb{P}, \mathbb{Q}$, define the Kullback-Leibler divergence by $D(\mathbb{P}||\mathbb{Q}) = \int \left(\log \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}$. The following result is Lemma 7 in [11]. It gives explicit formula for the Kullback-Leibler divergence between distributions generated by a special kind of covariance matrices.

Lemma 6.8. *For $i = 1, 2$, let $\Sigma_{(i)} = \begin{bmatrix} I_p & \lambda U_{(i)}V'_{(i)} \\ \lambda V_{(i)}U'_{(i)} & I_m \end{bmatrix}$ with $\lambda \in (0,1)$, $U_{(i)} \in O(p, r)$ and $V_{(i)} \in O(m, r)$. Let $\mathbb{P}_{(i)}$ denote the distribution of a random i.i.d. sample of size n from the $N_{p+m}(0, \Sigma_{(i)})$ distribution. Then*

$$D(\mathbb{P}_{(1)}||\mathbb{P}_{(2)}) = \frac{n\lambda^2}{2(1-\lambda^2)}\|U_{(1)}V'_{(1)} - U_{(2)}V'_{(2)}\|_{\text{F}}^2.$$

The main tool for our proof is Fano's lemma. The following version is adapted from [32, Lemma 3].

Proposition 6.1. *Let (Θ, ρ) be a metric space and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ a collection of probability measures. For any totally bounded $T \subset \Theta$, denote by $\mathcal{M}(T, \rho, \epsilon)$ the ϵ-packing number of T with respect to ρ, i.e., the maximal number of points in T whose pairwise minimum distance in ρ is at least ϵ. Define the Kullback-Leibler diameter of T by*

$$d_{\mathrm{KL}}(T) \triangleq \sup_{\theta, \theta' \in T} D(\mathbb{P}_\theta || \mathbb{P}_{\theta'}). \tag{51}$$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(\rho^2\big(\hat{\theta}(X), \theta\big) \geq \frac{\epsilon^2}{4} \right) \geq 1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)}. \tag{52}$$

Proof of Theorem 3.4. Due to the symmetry of the problem, we consider the loss $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$. The lower bound for the loss $\|P_{\widehat{V}} - P_V\|_{\mathrm{F}}^2$ has the same proof. The proof has three steps. In the first step, we derive the part $\frac{r s_u}{n\lambda_r^2}$ in the lower bound. In the second step, we derive the other part $\frac{s_u \log p}{n\lambda_r^2}$. Finally, we combine the two results in the third step.

Step 1. Let $U_0 = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \in O(p, r)$ and $V_0 = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \in O(m, r)$. For some $\epsilon_0 \in (0, \sqrt{r}]$ to be specified later, let

$$B(\epsilon_0) = \left\{ U \in O(p, r) : \mathrm{supp}(U) \subset [s_u], \|U - U_0\|_{\mathrm{F}} \leq \epsilon_0 \right\}.$$

and

$$T_0 = \left\{ \Sigma = \begin{bmatrix} I_p & \lambda_r U V_0' \\ \lambda_r V_0 U' & I_m \end{bmatrix} : U \in B(\epsilon_0) \right\}.$$

It is straightforward to verify that $T_0 \subset \mathcal{F}$. By Lemma 6.8,

$$\begin{aligned} d_{\mathrm{KL}}(T_0) &= \sup_{U_{(i)} \in B(\epsilon_0)} \frac{n\lambda_r^2}{2(1-\lambda_r^2)} \|U_{(1)} V_0' - U_{(2)} V_0'\|_{\mathrm{F}}^2 \\ &= \sup_{U_{(i)} \in B(\epsilon_0)} \frac{n\lambda_r^2}{2(1-\lambda_r^2)} \|U_{(1)} - U_{(2)}\|_{\mathrm{F}}^2 = \frac{2n\lambda_r^2 \epsilon_0^2}{1-\lambda_r^2}. \end{aligned} \tag{53}$$

Here, the second equality is due to the definition of V_0 and the third due to the definition of $B(\epsilon_0)$. We now establish a lower bound for the packing number of T_0. For some $\alpha \in (0,1)$ to be specified later, let $\{\widetilde{U}_{(1)}, \ldots, \widetilde{U}_{(N)}\} \subset O(p,r)$ be a maximal set such that $\text{supp}(\widetilde{U}_i) \subset [s_u]$, and for any $i \neq j \in [N]$,

$$\|\widetilde{U}_{(i)}\widetilde{U}'_{(i)} - U_0U'_0\|_{\text{F}} \leq \epsilon_0, \qquad \|\widetilde{U}_{(i)}\widetilde{U}'_{(i)} - \widetilde{U}_{(j)}\widetilde{U}'_{(j)}\|_{\text{F}} \geq \sqrt{2}\alpha\epsilon_0. \tag{54}$$

Then by [8, Lemma 1], for some absolute constant $C > 1$,

$$N \geq \left(\frac{1}{C\alpha}\right)^{r(s_u - r)}.$$

It is easy to see that the loss function $\|P_{U_{(i)}} - P_{U_{(j)}}\|_{\text{F}}^2$ on the subset T_0 equals $\|U_{(i)}U'_{(i)} - U_{(j)}U'_{(j)}\|_{\text{F}}^2$. Thus, for $\epsilon = \sqrt{2}\alpha\epsilon_0$ with sufficiently small α, $\log \mathcal{M}(T_0, \rho, \epsilon) \geq r(s_u - r)\log\frac{1}{C\alpha} \geq \frac{1}{2}rs_u \log\frac{1}{C\alpha}$. Taking $\epsilon_0^2 = c_1\frac{rs_u}{n\lambda_r^2}$ for sufficiently small c_1, we have

$$\inf_{\widehat{U}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\text{F}}^2 \geq \frac{\epsilon_0^2}{4}\right) \geq 1 - \frac{\frac{2c_1 rs_u}{1-\lambda_r^2} + \log 2}{\frac{1}{2}rs_u \log\frac{1}{C\alpha}}. \tag{55}$$

Since λ_r is bounded away from 1, we may choose sufficiently small c_0 and α, so that the right hand side of (55) can be lower bounded by 0.9. This completes the first step.

Step 2. The part $\frac{s_u \log p}{n\lambda_r^2}$ can be obtained from the rank-one argument spelled out in [9]. To be rigorous, consider the following subset of parameter space:

$$T_1 = \left\{\Sigma = \begin{bmatrix} I_p & \lambda_r U V_0' \\ \lambda_r V_0 U' & I_m \end{bmatrix} : U = \begin{bmatrix} I_{r-1} & 0 \\ 0 & u_r \end{bmatrix}, \right.$$
$$\left. u_r \in \mathbb{R}^{p-r+1}, ||u_r|| = 1, |\text{supp}(u_r)| \leq s_u - r + 1 \right\}.$$

Restricting on the set T_1, the loss function is

$$\|P_{U_{(i)}} - P_{U_{(j)}}\|_{\text{F}}^2 = \|u_{r,(i)}u'_{r,(i)} - u_{r,(j)}u'_{r,(j)}\|_{\text{F}}^2.$$

Let $X = [X_1 \ X_2]$ with $X_1 \in \mathbb{R}^{n\times(r-1)}$ and $X_2 \in \mathbb{R}^{n\times(p-r+1)}$, and $Y = [Y_1 \ Y_2]$ with $Y_1 \in \mathbb{R}^{n\times(r-1)}$ and $Y_2 \in \mathbb{R}^{n\times(m-r+1)}$. Then it is further equivalent to estimating u_1

under projection loss based on the observation (X_2, Y_2), because (X_2, Y_2) is a sufficient statistic for u_r. Applying the argument in [9, Appendix G] and choosing the appropriate constant, we have

$$\inf_{\widehat{U}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq C \frac{s_u \log p}{n\lambda_r^2} \wedge c_0 \right) \geq 0.9, \tag{56}$$

for some constant $C > 0$. This completes the second step.

Step 3. For any $\mathbb{P} \in \mathcal{P}$, by union bound, we have

$$\begin{aligned} & \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq \epsilon_1^2 \vee \epsilon_2^2\right) \\ \geq\ & 1 - \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 < \epsilon_1^2\right) - \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 < \epsilon_2^2\right) \\ =\ & \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq \epsilon_1^2\right) + \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq \epsilon_2^2\right) - 1. \end{aligned}$$

Taking $\sup_{\mathbb{P}\in\mathcal{P}}$ on both sides of the inequality, and letting $\epsilon_1^2 = C_1 \frac{r s_u}{n\lambda_r^2}$ in (55) and $\epsilon_2^2 = C_2 \frac{s_u \log p}{n\lambda_r^2} \wedge c_0$ in (56), we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq \epsilon_1^2 \vee \epsilon_2^2\right) \geq 0.9 + 0.9 - 1 = 0.8.$$

Thus, the proof is complete. □

7 Proofs of Technical Lemmas

In this section, we give proofs of the lemmas listed in Section 6. We first present an auxiliary result.

Lemma 7.1. *Assume* $\frac{1}{n}(s_u + s_v + \log(ep/s_u) + \log(em/s_v)) \leq C_1$ *for some constant* $c > 0$. *Then, for any* $C' > 0$, *there exists* $C > 0$ *only depending on* C' *such that*

$$\begin{aligned} \|U'\widehat{\Sigma}_x U - I\|_{\mathrm{op}} \vee \|(U'\widehat{\Sigma}_x U)^{1/2} - I\|_{\mathrm{op}} &\leq C\sqrt{\frac{1}{n}\left(s_u + \log\frac{ep}{s_u}\right)}, \\ \|V'\widehat{\Sigma}_y V - I\|_{\mathrm{op}} \vee \|(V'\widehat{\Sigma}_y V)^{1/2} - I\|_{\mathrm{op}} &\leq C\sqrt{\frac{1}{n}\left(s_v + \log\frac{em}{s_v}\right)}, \end{aligned}$$

with probability at least $1 - \exp(-C'(s_u + \log(ep/s_u))) - \exp(C'(s_v + \log(em/s_v)))$.

Proof. Using the definition of operator norm and the sparsity of U, we have

$$\|U'\widehat{\Sigma}_x U - I_r\|_{\mathrm{op}} = \|U'(\widehat{\Sigma}_x - \Sigma_x)U\|_{\mathrm{op}}$$
$$= \sup_{||v||=1} (Uv)'(\widehat{\Sigma}_x - \Sigma_x)(Uv) \leq \|U\|_{\mathrm{op}}^2 \|\widehat{\Sigma}_{xS_uS_u} - \Sigma_{xS_uS_u}\|_{\mathrm{op}},$$

where $\|U\|_{\mathrm{op}}^2 \leq \|\Sigma_x^{-1/2}\|_{\mathrm{op}}^2 \|\Sigma_x^{1/2} U\|_{\mathrm{op}}^2 \leq M$ and $\|\widehat{\Sigma}_{xS_uS_u} - \Sigma_{xS_uS_u}\|_{\mathrm{op}}$ is bounded by the desired rate with high probability according to Lemma 13 in [11]. Notice Lemma 13 in [11] was stated in the Gaussian case, but its proof also works for the sub-Gaussian case. Lemma 16 in [11] implies $\|(U'\widehat{\Sigma}_x U)^{1/2} - I\|_{\mathrm{op}} \leq C\|U'\widehat{\Sigma}_x U - I\|_{\mathrm{op}}$, and thus $\|(U'\widehat{\Sigma}_x U)^{1/2} - I\|_{\mathrm{op}}$ also shares same upper bound. The upper bound for $\|V'\widehat{\Sigma}_y V - I\|_{\mathrm{op}} \vee \|(V'\widehat{\Sigma}_y V)^{1/2} - I\|_{\mathrm{op}}$ can be derived by the same argument. Hence, the proof is complete. □

Proof of Lemma 6.3. Denote $F = [f_1, ..., f_r]$, $G = [g_1, ..., g_r]$ and $c_j = f_j' E b_j$. By $\|E\|_{\mathrm{op}} \leq 1$, we have $|c_j| \leq 1$. The left hand side of (25) is

$$\langle FKG', FG' - E\rangle = \langle K, I - F'EG\rangle = \sum_{l=1}^{r} k_{ll}(1 - c_l) \geq \min_{1\leq l\leq r} k_{ll} \sum_{l=1}^{r} (1 - c_l).$$

The right hand side of (25) is

$$\begin{aligned}
&\frac{\min_{1\leq l\leq r} k_{ll}}{2} \|FG' - E\|_{\mathrm{F}}^2 \\
&= \frac{\min_{1\leq l\leq r} k_{ll}}{2} \Big(\|FG'\|_{\mathrm{F}}^2 + \|E\|_{\mathrm{F}}^2 - 2\,\mathsf{Tr}(F'EG) \Big) \\
&\leq \frac{\min_{1\leq l\leq r} k_{ll}}{2} \Big(\mathsf{Tr}(F'FG'G) + \|E\|_{\mathrm{op}}\|E\|_* - 2\sum_{j=1}^{r} c_j \Big) \\
&\leq \min_{1\leq l\leq r} k_{ll} \sum_{j=1}^{r} (1 - c_j).
\end{aligned}$$

This completes the proof. □

Proof of Lemma 6.1. According to the definition (24), we have

$$\begin{aligned}
\|U-\widetilde{U}\|_{\rm op} &\leq \|U\|_{\rm op}\|(U'\widehat{\Sigma}_xU)^{1/2}-I\|_{\rm op}\|(U'\widehat{\Sigma}_xU)^{-1/2}\|_{\rm op}, \\
\|V-\widetilde{V}\|_{\rm op} &\leq \|V\|_{\rm op}\|(V'\widehat{\Sigma}_yV)^{1/2}-I\|_{\rm op}\|(V'\widehat{\Sigma}_yV)^{-1/2}\|_{\rm op}, \\
\|\widetilde{\Lambda}-\Lambda\|_{\rm op} &\leq \|(U'\widehat{\Sigma}_xU)^{1/2}-I\|_{\rm op}\|\Lambda(V'\widehat{\Sigma}_yV)^{1/2}\|_{\rm op} \\
&\qquad +\|\Lambda\|_{\rm op}\|(V'\widehat{\Sigma}_yV)^{1/2}-I\|_{\rm op}, \\
\|\widetilde{A}-A\|_{\rm op} &\leq \|U\|_{\rm op}\|V-\widetilde{V}\|_{\rm op}+\|\widetilde{V}\|_{\rm op}\|U-\widetilde{U}\|_{\rm op}.
\end{aligned}$$

Applying Lemma 7.1, the proof is complete. □

Proof of Lemma 6.2. By the definition of $\widetilde{U}$, we have $\widetilde{U}'\widehat{\Sigma}_x\widetilde{U}=I$, and thus $\widehat{\Sigma}_x^{1/2}\widetilde{U}\in O(p,r)$. Similarly $\widehat{\Sigma}_y^{1/2}\widetilde{V}\in O(m,r)$. Thus,

$$\|\widehat{\Sigma}_x^{1/2}\widetilde{A}\widehat{\Sigma}_y^{1/2}\|_{\rm op}\leq\|\widehat{\Sigma}_x^{1/2}\widetilde{U}\|_{\rm op}\|\widehat{\Sigma}_y^{1/2}\widetilde{V}\|_{\rm op}\leq 1. \tag{57}$$

Now let us use the notation $Q=\widehat{\Sigma}_x^{1/2}\widetilde{A}\widehat{\Sigma}_y^{1/2}$. Then, by the definition of $\widetilde{A}$, we have $Q'Q=\widehat{\Sigma}_y^{1/2}V(V'\widehat{\Sigma}_yV)^{-1}V'\widehat{\Sigma}_y^{1/2}$, and

$$\mathsf{Tr}(Q'Q)=\mathsf{Tr}((V'\widehat{\Sigma}_yV)^{-1}(V'\widehat{\Sigma}_yV))=r. \tag{58}$$

Combining (57) and (58), it is easy to see that all eigenvalues of $Q'Q$ are 1. Thus, we have $\|Q\|_*=r$ and $\|Q\|_{\rm op}=1$. The proof is complete. □

Proof of Lemma 6.4. Using triangle inequality, $||\widehat{\Sigma}_{xy}-\widetilde{\Sigma}_{xy}||_\infty$ can be upper bounded by the following sum,

$$\begin{aligned}
&||\widehat{\Sigma}_{xy}-\Sigma_{xy}||_\infty+||(\widehat{\Sigma}_x-\Sigma_x)U\Lambda V'\Sigma_y||_\infty \\
&+||\Sigma_xU\Lambda V'(\widehat{\Sigma}_y-\Sigma_y)||_\infty+||(\widehat{\Sigma}_x-\Sigma_x)U\Lambda V'(\widehat{\Sigma}_y-\Sigma_y)||_\infty.
\end{aligned}$$

The first term can be bounded by the desired rate by union bound and Bernstein's inequality [25, Prop. 5.16]. For the second term, it can be written as

$$\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n(X_{ij}[X_i'U\Lambda V'\Sigma_y]_k-\mathbb{E}X_{ij}[X_i'U\Lambda V'\Sigma_y]_k)\right|,$$

where X_{ij} is the j-th element of X_i and the notation $[\cdot]_k$ means the k-th element of the referred vector. Thus, it is a maximum over average of centered sub-exponential random variables. Then, by Bernstein's inequality and union bound, it is also bounded by the desired rate. Similarly, we can bound the third term. For the last term, it can be bounded by $\sum_{l=1}^r \lambda_l ||(\widehat{\Sigma}_x - \Sigma_x)u_l v_l'(\widehat{\Sigma}_y - \Sigma_y)||_\infty$, where for each l, $||(\widehat{\Sigma}_x - \Sigma_x)u_l v_l'(\widehat{\Sigma}_y - \Sigma_y)||_\infty$ can be written as

$$\max_{j,k} \left| \left(\frac{1}{n} \sum_{i=1}^n (X_{ij} X_i' u_l - \mathbb{E} X_{ij} X_i' u_l) \right) \left(\frac{1}{n} \sum_{i=1}^n (Y_{ik} Y_i' v_l - \mathbb{E} Y_{ik} Y_i' v_l) \right) \right|.$$

It can be bounded by the rate $\frac{\log(p+m)}{n}$ with the desired probability using union bound and Bernstein's inequality. Hence, the last term can be bounded by $\frac{\lambda_1 r \log(p+m)}{n}$. Under the assumption that $r\sqrt{\frac{\log(p+m)}{n}}$ is bounded by a constant, it can further be bounded by the rate $\sqrt{\frac{\log(p+m)}{n}}$ with high probability. Combining the bounds of the four terms, the proof is complete. □

Proof of Lemma 6.6. By the definition of U^*, we have $\Sigma_{xy} V^{(0)} = \Sigma_x U^*$. Thus,

$$\max_{1 \le j \le p} ||[\widehat{\Sigma}_{xy} V^{(0)} - \widehat{\Sigma}_x U^*]_{j\cdot}|| \le \max_{1 \le j \le p} ||[(\widehat{\Sigma}_{xy} - \Sigma_{xy}) V^{(0)}]_{j\cdot}|| + \max_{1 \le j \le p} ||[(\widehat{\Sigma}_x - \Sigma_x) U^*]_{j\cdot}||.$$

Let us first bound $\max_{1 \le j \le p} ||[(\widehat{\Sigma}_x - \Sigma_x) U^*]_{j\cdot}||$. Note that the sample covariance can be written as

$$\widehat{\Sigma}_x = \Sigma_x^{1/2} \left(\frac{1}{n} \sum_{i=1}^n Z_i Z_i' \right) \Sigma_x^{1/2},$$

where $\{Z_i\}_{i=1}^n$ are i.i.d. sub-Gaussian vectors with $||Z_i||_{\psi_2} = 1$. Let T_j' be the j-th row of $\Sigma_x^{1/2}$, and then we have

$$[(\widehat{\Sigma}_x - \Sigma_x) U^*]_{j\cdot} = \frac{1}{n} \sum_{i=1}^n (T_j' Z_i Z_i' \Sigma_x^{1/2} U^* - T_j' \Sigma_x^{1/2} U^*).$$

For each i and j, define vector

$$W_i^{(j)} = \begin{bmatrix} T_j' Z_i \\ (U^*)' \Sigma_x^{1/2} Z_i \end{bmatrix}.$$

Since $T_j' Z_i Z_i' \Sigma_x^{1/2} U^*$ is a submatrix of $W_i^{(j)}(W_i^{(j)})'$, we have

$$|||[(\widehat{\Sigma}_x - \Sigma_x)U^*]_{j\cdot}|| \leq \|\frac{1}{n}\sum_{i=1}^n (W_i^{(j)}(W_i^{(j)})' - \mathbb{E}W_i^{(j)}(W_i^{(j)})')\|_{\mathrm{op}}.$$

Hence, for any $t > 0$, we have

$$\begin{aligned}
& \mathbb{P}\left\{\max_{1\leq j\leq p} |||[(\widehat{\Sigma}_x - \Sigma_x)U^*]_{j\cdot}|| > t\right\} \\
\leq & \sum_{j=1}^p \mathbb{P}\left\{\|\frac{1}{n}\sum_{i=1}^n (W_i^{(j)}(W_i^{(j)})' - \mathbb{E}W_i^{(j)}(W_i^{(j)})')\|_{\mathrm{op}} > t\right\} \\
\leq & \sum_{j=1}^p \exp\left(C_1 r - C_2 n \min\left\{\frac{t}{\|\mathcal{W}^{(j)}\|_{\mathrm{op}}}, \frac{t^2}{\|\mathcal{W}^{(j)}\|_{\mathrm{op}}^2}\right\}\right), \qquad (59)
\end{aligned}$$

where $\mathcal{W}^{(j)} = \mathbb{E}W_i^{(j)}(W_i^{(j)})'$, and we have used concentration inequality for sample covariance [25, Thm. 5.39]. Since $\|\mathcal{W}^{(j)}\|_{\mathrm{op}} \leq C_3$ for some constant C_3 only depending on M, (59) can be bounded by

$$\exp\left(C_1'(r + \log p) - C_2' n(t \wedge t^2)\right).$$

Take $t^2 = C_4 \frac{r+\log p}{n}$ for some sufficiently large C_4, and under the assumption (23), $\max_{1\leq j\leq p} |||[(\widehat{\Sigma}_x - \Sigma_x)U^*]_{j\cdot}|| \leq C\sqrt{\frac{r+\log p}{n}}$ with probability at least $1 - \exp(-C'(r + \log p))$. Similar arguments lead to the bound of $\max_{1\leq j\leq p} |||[(\widehat{\Sigma}_{xy} - \Sigma_{xy})V^{(0)}]_{j\cdot}||$. Let us sketch the proof. Note that we may write

$$[(\widehat{\Sigma}_{xy} - \Sigma_{xy})V^{(0)}]_j = \frac{1}{n}\sum_{i=1}^n \left(T_j' Z_i Y_i' V^{(0)} - \mathbb{E}(T_j' Z_i Y_i' V^{(0)})\right).$$

Then, define

$$H_i^{(j)} = \begin{bmatrix} T_j' Z_i \\ (V^{(0)})' Y_i \end{bmatrix},$$

and we have

$$\max_{1\leq j\leq p} |||[(\widehat{\Sigma}_{xy} - \Sigma_{xy})V^{(0)}]_j|| \leq \max_{1\leq j\leq p} \|\frac{1}{n}\sum_{i=1}^n (H_i^{(j)}(H_i^{(j)})' - \mathbb{E}H_i^{(j)}(H_i^{(j)})')\|_{\mathrm{op}}.$$

Using the same argument, we can bound this term by $C\sqrt{\frac{r+\log p}{n}}$ with probability at least $1 - \exp(-C'(r + \log p))$. Thus, the proof is complete. □

References

[1] T. W. Anderson. An introduction to multivariate statistical analysis. 1958.

[2] F. Andre, B. Job, P. Dessen, A. Tordai, S. Michiels, C. Liedtke, C. Richon, K. Yan, B. Wang, and G. Vassal. Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array. *Clinical Cancer Research*, 15(2):441–451, 2009.

[3] B. B. Avants, P. A. Cook, L. Ungar, J. C. Gee, and M. Grossman. 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, 2010.

[4] S. R. Becker, E. J. Candès, and M. C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[5] F. Bertucci, P. Finetti, J. Rougemont, E. Charafe-Jauffret, V. Nasser, B. Loriod, J. Camerlo, R. Tagett, C. Tarpin, and G. Houvenaeghel. Gene expression profiling for molecular characterization of inflammatory breast cancer and prediction of response to chemotherapy. *Cancer Research*, 64(23):8558–8565, 2004.

[6] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[8] T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. *The Annals of Statistics (in press)*, 2013.

[9] M. Chen, C. Gao, Z. Ren, and H. H. Zhou. Sparse CCA via precision adjusted iterative thresholding. *arXiv preprint arXiv:1311.6186*, 2013.

[10] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3): 434–448, 2007.

[11] C. Gao, Z. Ma, Z. Ren, and H. H. Zhou. Minimax estimation in sparse canonical correlation analysis. *arXiv preprint arXiv:1405.1595*, 2014.

[12] G. H. Golub and C. F. Van Loan. Matrix computations (3rd ed.). 1996.

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

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

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

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

[17] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, 121(7):2750–2767, 2011.

[18] X.-J. Ma, S. Dahiya, E. Richardson, M. Erlander, and D. C. Sgroi. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Research*, 11(1):R7, 2009.

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

[20] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[21] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

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

[23] G.W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990.

[24] A. W. van der Vaart and J. A. Wellner. *Weak Convergence*. Springer, 1996.

[25] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[26] V. Q. Vu, J. Cho, J. Lei, and K. Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems*, pages 2670–2678, 2013.

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

[28] Y. X. R. Wang, K. Jiang, L. J. Feldman, P. J. Bickel, and H. Huang. Inferring gene association networks using sparse canonical correlation analysis. *arXiv preprint arXiv:1401.6504*, 2014.

[29] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

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

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

[32] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

[33] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.