Sparse CCA: Adaptive Estimation and Computational Barriers *

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

[1] *Yale University*
[2] *University of Pennsylvania*

Abstract

Canonical correlation analysis (CCA) is a classical and important multivariate technique for exploring the relationship between two sets of variables. It has applications in many fields including genomics and imaging, to extract meaningful features as well as to use the features for subsequent analysis. This paper considers adaptive and computationally tractable estimation of leading sparse canonical directions when the ambient dimensions are high. Three intrinsically related problems are studied to fully address the topic. First, we establish the minimax rates of the problem under prediction loss. Separate minimax rates are obtained for canonical directions of each set of random variables under mild conditions. There is no structural assumption needed on the marginal covariance matrices as long as they are well conditioned. Second, we propose a computationally feasible two-stage estimation procedure, which consists of a convex programming based initialization stage and a group-Lasso based refinement stage, to attain the minimax rates under an additional sample size condition. Finally, we provide evidence that the additional sample size condition is essentially necessary for any randomized polynomial-time estimator to be consistent, assuming hardness of the Planted Clique detection problem. The computational lower bound is faithful to the Gaussian models used in the paper, which is achieved by a novel construction of the reduction scheme and an asymptotic equivalence theory for Gaussian discretization that is necessary for computational complexity to be well-defined. As a byproduct, we also obtain computational lower bound for the sparse PCA problem under the Gaussian spiked covariance model. This bridges a gap in the sparse PCA literature.

Keywords. Convex programming, group-Lasso, Minimax rates, Computational complexity, Planted Clique, Sparse CCA (SCCA), Sparse PCA (SPCA)

1 Introduction

Canonical correlation analysis (CCA) [20] is one of the most classical and important tools in multivariate statistics [1, 27]. For two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^m$, at the population

---

level, CCA finds successive vectors $u_j \in \mathbb{R}^p$ and $v_j \in \mathbb{R}^m$ (called *canonical directions*) that solve

$$\max_{a,b} \quad a'\Sigma_{xy}b,$$

$$\text{subject to} \quad a'\Sigma_x a = b'\Sigma_y b = 1, \ a'\Sigma_x u_l = b'\Sigma_y v_l = 0, \ \forall 0 \le l \le j-1, \tag{1}$$

where $\Sigma_x = \text{Cov}(X), \Sigma_y = \text{Cov}(Y), \Sigma_{xy} = \text{Cov}(X, Y), u_0 = 0$, and $v_0 = 0$. Since our primary interest lies in the covariance structure among X and Y, we assume that their means are zeros from here on. Then the linear combinations $(u_j'X, v_j'Y)$ are the j-th *canonical variates*. This technique has been widely used in various scientific fields to explore the relation between two sets of variables. In practice, one does not have knowledge about the population covariance, and Σ_x, Σ_y, and Σ_{xy} are replaced by their sample versions $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$, and $\widehat{\Sigma}_{xy}$ in (1).

Recently, there have been growing interests in applying CCA to analyzing high-dimensional datasets, where the dimensions p and m could be much larger than the sample size n. It has by now been well understood that classical CCA breaks down in this regime [22, 4, 16]. Motivated by genomics, neuroimaging and other applications, people have become interested in seeking sparse leading canonical direction vectors. Various estimation procedures imposing sparsity on canonical directions have been developed in the literature, which are usually termed *sparse CCA*. See, for example, [37, 38, 28, 19, 24, 32, 3]. In addition to its use as a high-dimensional multivariate analysis tool, sparse CCA is also used to extract meaningful features in data for subsequent analysis. For example, Wang et al. [34] proposed to employ sparse CCA to compute edge weights in gene networks and then to infer gene relationships by community detection on the constructed networks.

The theoretical aspect of sparse CCA has also been investigated in the literature. A useful model for studying sparse CCA is the *canonical pair model* proposed in [12]. In particular, suppose there are r pairs of canonical directions (variates) among the two sets of variables, then the canonical pair model reparameterizes the cross-covariance matrix as

$$\Sigma_{xy} = \Sigma_x U \Lambda V' \Sigma_y, \quad \text{where} \quad U'\Sigma_x U = V'\Sigma_y V = I_r. \tag{2}$$

Here $U = [u_1, ..., u_r]$ and $V = [v_1, ..., v_r]$ collect the canonical direction vectors and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ with $1 > \lambda_1 \ge \cdots \ge \lambda_r > 0$ are the ordered canonical correlations. Let $S_u = \text{supp}(U)$ and $S_v = \text{supp}(V)$ be the indices of nonzero rows of U and V. One way to impose sparsity on the canonical directions is to require the sizes of S_u and S_v to be small, namely $|S_u| \le s_u$ and $|S_v| \le s_v$ for some $s_u \le p$ and $s_v \le m$. Under this model, Gao et al. [16] showed that the minimax rate for estimating U and V under the joint loss function $\|\widehat{U}\widehat{V}' - UV'\|_F^2$ is

$$\frac{1}{n\lambda_r^2} \left(r(s_u + s_v) + s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v} \right). \tag{3}$$

However, to achieve the rate, Gao et al. [16] used a computationally infeasible and non-adaptive procedure, which requires exhaustive search of all possible subsets with the given cardinality and the knowledge of s_u and s_v. Moreover, it is unclear from (3) whether the estimation of U per se interferes with that of V and vice versa.

2

The goal of the present paper is to study three fundamental and practically important questions: (1) What are the minimax rates for estimating the canonical directions on the two sets of variables separately? (2) Is there a computationally efficient and sparsity-adaptive method that achieves the optimal rates? (3) What is the price one has to pay to achieve the optimal rates in a computationally efficient way?

Under the canonical pair model and Gaussianity assumption, we first characterize the separate minimax rates for estimating U and V under a natural prediction loss function. Moreover, we provide an affirmative answer to the second question by proposing a two-stage estimation procedure where both stages are based on convex programming which admit efficient computation. The resulting estimator is shown to achieve the minimax rates under an extra sample size condition. Importantly, both the minimax characterization and the adaptive procedure require no structural assumption on the marginal covariance matrices Σ_x and Σ_y other than them being well-conditioned. To the best of our limited knowledge, this is the *first* computationally feasible sparse CCA algorithm that achieves optimal statistical performance without imposing restrictive assumptions on Σ_x and Σ_y. Last but not least, we provide a computational lower bound to show that the additional sample size condition is essentially the price one has to pay in order to achieve consistency, a weaker requirement than minimax optimality whenever the minimax rates converge to zero, by any computationally efficient algorithm. Our computational lower bound is faithful to the Gaussian canonical pair model used in the paper, which distinguishes itself from previous results on the related sparse PCA problem [6, 33] which required generalization to much larger space of distributions. In fact, as a byproduct of our arguments, we also obtain a computational lower bound for the sparse PCA problem under the Gaussian spiked covariance model [21].

1.1 Main contributions

We introduce in more detail the main contributions of the present paper from three different viewpoints as suggested by the three questions we raised above.

Separate minimax rates The joint loss $\|\widehat{U}\widehat{V}' - UV'\|_{\mathrm{F}}^2$ studied by [16] characterizes the joint estimation error of both canonical directions U and V. In this paper, we provide a finer analysis by studying individual estimation errors of U and V separately under a natural loss function that can be interpreted as prediction error of canonical variates. The exact definition of the loss functions is given in Section 2. Separate minimax rates are obtained for U and V. In particular, we show that the error in estimating U depends only n, r, λ_r, p and s_u, but not on either m or s_v. Consequently, if U is sparser than V, then convergence rate for estimating U can be faster than that for estimating V. Such a difference is not reflected by the joint loss, since its minimax rate (3) is determined by the slower between the rates of estimating U and V.

Adaptive estimation To achieve optimal rates adaptively, we propose a computationally efficient algorithm under the canonical pair model. The algorithm is a two-stage estimation

procedure. In the first stage, we propose a convex programming for sparse CCA based on a tight convex relaxation of a combinatorial program in [16] by considering the smallest convex set containing all matrices of the form AB' with both A and B being rank-r orthogonal matrices. The convex programming can be efficiently solved by the Alternating Direction Method with Multipliers (ADMM) algorithm [14, 10]. Based on the output of the first stage, we formulate a sparse linear regression problem in the second stage to improve rates of convergence, and the final estimator $\widehat{U}$ and $\widehat{V}$ can be obtained via a group-Lasso algorithm [40]. Under the sample size condition that

$$n \geq C \frac{s_u s_v \log(p+m)}{\lambda_r^2}, \tag{4}$$

for some sufficiently large constant $C > 0$, we show $\widehat{U}$ and $\widehat{V}$ recover the true canonical directions U and V within optimal error rates adaptively with high probability. It is worthwhile to point out that a naive application of Lasso algorithm leads to an inferior rate.

As was pointed out in [12] and [16], sparse CCA is a more involved problem than the well-studied sparse PCA. A naive application of sparse PCA algorithm to sparse CCA leads to possibly inconsistent results, as was shown in [12]. The additional difficulty in sparse CCA comes from two sources. First, due to the presence of the nuisance parameters Σ_x and Σ_y, the cross-covariance Σ_{xy} is not a sparse matrix itself. This makes the subset selection procedure in [21] and [11] inapplicable in sparse CCA when Σ_x and Σ_y are unknown. Second, the sparse canonical direction matrices U and V do not have orthogonal columns in the usual Euclidean metric. Instead, their columns are orthogonal with respect to Σ_x and Σ_y, which are unknown objects that have to be estimated from data. In high dimensional settings, estimators of Σ_x and Σ_y can be inconsistent without strong structural assumptions imposed. These difficulties were overcome by a computationally infeasible procedure in [16]. The key observation in [16] was that the sample covariance matrix, when restricted on a subset of variables of true sparsity size, is a good estimator of the true sub covariance matrix under operator norm (Lemma 14 in [16]), though the whole sample covariance matrix may not be consistent in high-dimensions. However, this required the algorithm to conduct an exhaustive search over all possible subsets, which is computationally infeasible.

In the present paper, computational intractability is further overcome by the proposed convex relaxation which is not only tight but also preserves the desired curvature of the problem by establishing the corresponding curvature lemma (Lemma 6.3 below). The lemma can be viewed as a convex extension of the generalized sin-theta theorem established in [16]. Together with the curvature lemma, we show that the error matrix of the convex optimization lies in a *generalized cone*, a concept that we coin to extend the well-known cone condition in sparse linear regression [8]. A restricted eigenvalue property is established on the generalized cone, which leads to the desired convergence rate of our proposed estimator.

Computational lower bound   As cost of computational feasibility, we require the sample size condition (4) for the adaptive procedure to achieve optimal rates of convergence. Assuming hardness of the Planted Clique detection problem, we provide a computational lower

4

bound to show that a condition of this kind is unavoidable for any computationally feasible estimation procedure to achieve consistency. To rigorously establish the computational lower bound, we adopt the framework of asymptotically equivalent discretized model developed in [26]. Up to an asymptotically equivalent discretization which is necessary for computational complexity to be well-defined, our computational lower bound is established directly for the Gaussian canonical pair model used throughout the paper.

An analogous sample size condition, namely $n \geq Cs^2 \log p/\lambda^2$ where s is the sparsity of the leading eigenvector and λ the gap between the leading eigenvalue and the rest of the spectrum, has been imposed in the sparse PCA literature (see [21, 25, 11, 31, 7]). In an important paper, Berthet and Rigollet [6] showed that if there existed a polynomial-time algorithm for a generalized sparse PCA detection problem while the condition is violated, then the algorithm could be made (in randomized polynomial-time) into a detection method for the Planted Clique problem in a regime where it is believed to be computationally intractable. However, both the null and the alternative hypotheses in the sparse PCA detection problem were generalized to include all multivariate distributions whose quadratic forms satisfy certain uniform tail probability bounds. The same remark also applies to the subsequent work on sparse PCA estimation [33]. Hence, the computational lower bound in sparse PCA was only established for such enlarged parameter spaces. As a byproduct of our analysis, we establish the desired computational lower bound for sparse PCA in the (discretized) Gaussian spiked covariance model. This strengthens computational lower bounds in [6, 33] and bridges the gap between the computational lower bound and the minimax/adaptive estimation for the Gaussian sparse PCA.

1.2 Organization

After the introduction of notation below, the rest of the paper is organized as follows. In Section 2, we formulate the sparse CCA problem by defining its parameter space and loss function. Section 3 presents minimax rates of the problem by introducing a rate-optimal estimator and establishing the corresponding minimax lower bound. Section 4 is devoted to adaptive estimation, where we propose a two-stage estimator via a novel form of convex relaxation. The proposed estimator is shown to be minimax optimal under an additional sample size condition. The condition is shown to be essentially necessary for all randomized polynomial-time estimator in Section 5. Section 6 presents proofs of theoretical results in Section 4. Due to page limits, computational lower bounds on sparse PCA, implementation of the adaptive procedure, numerical studies and additional proofs are all deferred to the supplement.

1.3 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 and $\lfloor a \rfloor$ the largest integer no

larger 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 $A_i.$. For any subset $J \subset [p] \times [k]$ of indices, we use $A_J = (a_{ij} \mathbf{1}_{\{(i,j) \in J\}})$ to denote the $p \times k$ matrix whose entries on J are the same as those in A and the entries outside J are all zeros. 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 $A_{J_1 \times [k]} \in \mathbb{R}^{p \times k}$ while $A_{J_1.}$ stands for the corresponding nonzero submatrix which is of size $|J_1| \times k$. For any square matrix $A = (a_{ij})$, denote its trace by $\mathsf{Tr}(A) = \sum_i a_{ii}$. For two square matrices A and B, the relation $A \preceq B$ means $B - A$ is positive semidefinite. Moreover, let $O(p, k) = \{A \in \mathbb{R}^{p \times k} : A'A = I_k\}$ denote the set of all $p \times k$ orthogonal matrices and $O(k) = O(k, k)$. For any matrix $A \in \mathbb{R}^{p \times k}$, P_A stands for the $p \times p$ projection matrix onto the column space of A. The notation $\sigma_i(A)$ stands for its i-th largest singular value. In particular, $\sigma_{\max}(A) = \sigma_1(A)$ and $\sigma_{\min}(A) = \sigma_{p \wedge k}(A)$. 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 $\mathrm{supp}(A) = \{i \in [p] : ||A_i.|| > 0\}$, the index set of nonzero rows. 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 $\langle A, B \rangle = \mathsf{Tr}(A'B)$. For two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, the total variation distance is defined as $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) = \sup_B |\mathbb{P}(B) - \mathbb{Q}(B)|$. We also write $\mathsf{TV}(p, q)$ if p and q are the densities of $\mathbb{P}$ and $\mathbb{Q}$, respectively. Given a random element X, $\mathcal{L}(X)$ denotes its probability distribution. 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. Notation $\mathbb{P}$ and $\mathbb{E}$ stand for generic probability and expectation when the distribution is clear from the context.

2 Problem Formulation

2.1 Parameter space

Consider a canonical pair model where the observed pairs of measurement vectors (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d. from a multivariate Gaussian distribution $N_{p+m}(0, \Sigma)$ where

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix},$$

with the cross-covariance matrix Σ_{xy} satisfying (2). We are interested in the situation where the leading canonical directions are specified by sparse vectors. One way to quantify the level of sparsity is to bound how many nonzero rows there are in the U and V matrices. This notion of sparsity has been used previously in both sparse PCA [11, 31] and sparse CCA [16] problems when one seeks multiple sparse vectors simultaneously.

Recall that for any matrix A, $\mathrm{supp}(A)$ collects the indices of nonzero rows in A. Adopting the above notion of sparsity, we define $\mathcal{F}(s_u, s_v, p, m, r, \lambda; M)$ to be the collection of all

covariance matrices Σ with the structure (2) satisfying

1. $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{m \times r}$ with $|\operatorname{supp}(U)| \leq s_u$ and $|\operatorname{supp}(V)| \leq s_v$;
2. $\sigma_{\min}(\Sigma_x) \wedge \sigma_{\min}(\Sigma_y) \geq M^{-1}$ and $\sigma_{\max}(\Sigma_x) \vee \sigma_{\max}(\Sigma_y) \leq M$; $\qquad$ (5)
3. $\lambda_r \geq \lambda$ and $\lambda_1 \leq 1 - M^{-1}$.

The probability space we consider is

$$\mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M) = \Big\{ \mathcal{L}(X_1, ..., X_n) : X_i \overset{iid}{\sim} N_{p+m}(0, \Sigma)$$
$$\text{with } \Sigma \in \mathcal{F}(s_u, s_v, p, m, r, \lambda; M) \Big\}, \qquad (6)$$

where n is the sample size. We shall allow $s_u, s_v, p, m, r, \lambda$ to vary with n, while $M > 1$ is restricted to be an absolute constant.

Our goal is to achieve the optimal rates of convergence adaptively on a large collection of parameter spaces of the above form. To this end, we need to further specify the loss function we use, to which we now turn.

2.2 Prediction loss

From now on, the presentation of definitions and results will focus on U only since those for V can be obtained via symmetry. Given an estimator $\widehat{U} = [\widehat{u}_1, \ldots, \widehat{u}_r]$ of the leading canonical directions for X, a natural way of assessing its quality is to see how well it predicts the values of the canonical variables $U'X^\star \in \mathbb{R}^r$ for a new observation $X^\star$ which is independent of and identically distributed as the training sample used to obtain $\widehat{U}$. This leads us to consider the following loss function

$$L(\widehat{U}, U) = \inf_{W \in O(r)} \mathbb{E}^\star \| W' \widehat{U}' X^\star - U' X^\star \|^2, \qquad (7)$$

where $\mathbb{E}^\star$ means taking expectation only over $X^\star$ and so $L(\widehat{U}, U)$ is still a random quantity due to the randomness of $\widehat{U}$. Since $L(\widehat{U}, U)$ is the expected squared error for predicting the canonical variables $U'X^\star$ via $\widehat{U}'X^\star$, we refer to it as prediction loss from now on. It is worth noting that the introduction of an $r \times r$ orthogonal matrix W is unavoidable. To see this, we can simply consider the case where $\lambda_1 = \cdots = \lambda_r = \lambda$ in (2), then we can replace the pair (U, V) in (2) by (UW, VW) for any $W \in O(r)$. In other words, the canonical directions are only determined up to a joint orthogonal transform. If we work out the $\mathbb{E}^\star$ part in the definition (7), then the loss function can be equivalently defined as

$$L(\widehat{U}, U) = \inf_{W \in O(r)} \mathsf{Tr}[(\widehat{U}W - U)' \Sigma_x (\widehat{U}W - U)]. \qquad (8)$$

By symmetry, we can define $L(\widehat{V}, V)$ by simply replacing U, $\widehat{U}$, $X^\star$ and Σ_x in (7) and (8) with V, $\widehat{V}$, $Y^\star$ and Σ_y.

A related loss function is the squared subspace distance $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$. By Proposition 9.2 in the supplementary material, the prediction loss $L(\widehat{U}, U)$ is a stronger loss function than the squared subspace distance. That is, $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \leq CL(\widehat{U}, U)$ for some constant $C > 0$ only depending on M.

3 Minimax Rates

To provide a benchmark for any estimation procedure, we determine the minimax rates of the statistical problem formulated in the previous section. To this end, we first provide a minimax upper bound using a combinatorial optimization procedure, and then show that the resulting rate is optimal by further providing a matching minimax lower bound.

Let $(X_i', Y_i')' \in \mathbb{R}^{p+m}, i = 1, \ldots, n$, be i.i.d. observations following $N_{p+m}(0, \Sigma)$ for some $\Sigma \in \mathcal{F}(s_u, s_v, p, m, r, \lambda; M)$. For notational convenience, we assume the sample size is divisible by three, i.e., $n = 3n_0$ for some $n_0 \in \mathbb{N}$.

Procedure To obtain minimax upper bound, we propose a two-stage combinatorial optimization procedure. We split the data into three equal size batches $\mathcal{D}_0 = \{(X_i', Y_i')'\}_{i=1}^{n_0}$, $\mathcal{D}_1 = \{(X_i', Y_i')'\}_{i=n_0+1}^{2n_0}$ and $\mathcal{D}_2 = \{(X_i', Y_i')'\}_{i=2n_0+1}^{n}$, and denote the sample covariance matrices computed on each batch by $\widehat{\Sigma}_x^{(j)}, \widehat{\Sigma}_y^{(j)}$ and $\widehat{\Sigma}_{xy}^{(j)}$ for $j \in \{0, 1, 2\}$.

In the first stage, we find $(\widehat{U}^{(0)}, \widehat{V}^{(0)})$ which solves the following program:

$$
\max_{L \in \mathbb{R}^{p \times r}, R \in \mathbb{R}^{m \times r}} \quad \mathsf{Tr}(L'\widehat{\Sigma}_{xy}^{(0)}R),
$$
$$
\text{subject to} \quad L'\widehat{\Sigma}_x^{(0)}L = R'\widehat{\Sigma}_y^{(0)}R = I_r, \text{ and} \tag{9}
$$
$$
|\mathrm{supp}(L)| \le s_u, |\mathrm{supp}(R)| \le s_v.
$$

In the second stage, we further refine the estimator for U by finding $\widehat{U}^{(1)}$ solving

$$
\min_{L \in \mathbb{R}^{p \times r}} \quad \mathsf{Tr}(L'\widehat{\Sigma}_x^{(1)}L) - 2\,\mathsf{Tr}(L'\widehat{\Sigma}_{xy}^{(1)}\widehat{V}^{(0)})
$$
$$
\text{subject to} \quad |\mathrm{supp}(L)| \le s_u. \tag{10}
$$

The final estimator is a normalized version of $\widehat{U}^{(1)}$, defined as

$$
\widehat{U} = \widehat{U}^{(1)}((\widehat{U}^{(1)})'\widehat{\Sigma}_x^{(2)}\widehat{U}^{(1)})^{-1/2}. \tag{11}
$$

The motivation of the second stage will be discussed below after the statement of Theorem 3.1. We remark that the purpose of sample splitting employed in the above procedure is to facilitate the proof.

Theory and discussion We now state the bounds related to the initial and final estimators together with discussion on the intuition behind the proposed procedure. The first upper bound concerns the initial estimator $(\widehat{U}^{(0)}, \widehat{V}^{(0)})$.

Theorem 3.1. *Assume*

$$
\frac{1}{n}\left(r(s_u + s_v) + s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v}\right) \le c \tag{12}
$$

for some sufficiently small $c > 0$. Then for any $C' > 0$, there exists $C > 0$ only depending on C' such that

$$
\|\Sigma_x^{1/2}\left(\widehat{U}^{(0)}(\widehat{V}^{(0)})' - UV'\right)\Sigma_y^{1/2}\|_{\mathrm{F}}^2 \le \frac{C}{n\lambda^2}\left(r(s_u + s_v) + s_u \log \frac{ep}{s_u} + s_v \log \frac{em}{s_v}\right),
$$

8

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

The program (9) was first proposed in [16] as a sparsity constrained version of the classical CCA formulation. Theorem 3.1 can then be viewed as a special case of Theorem 1 in [16]. We nonetheless present its proof in Section 9.1 in the supplementary material for the paper to be self-contained. Using Wedin's sin-theta theorem, one can directly derive from the bound in Theorem 3.1 an upper bound for estimating U under the prediction loss (7). However, the resulting bound will then involve the sparsity level s_v and the ambient dimension m of the V matrix, which is sub-optimal. The second stage in the procedure is thus proposed to further pursue the optimal estimation rates, as is shown in the following theorem.

Theorem 3.2. *Assume* (12) *holds for some sufficiently small* $c > 0$. *Then for any* $C' > 0$, *there exists* $C > 0$ *only depending on* C' *such that*

$$L(\widehat{U}, U) \leq \frac{C}{n\lambda^2} s_u \left(r + \log \frac{ep}{s_u} \right),\tag{13}$$

with $\mathbb{P}$-*probability at least* $1 - \exp\left(-C'(s_u + \log(ep/s_u))\right) - \exp\left(-C'(s_v + \log(em/s_v))\right)$ *uniformly over* $\mathbb{P} \in \mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M)$.

The motivation for the second stage is as follows. First, if we were given the knowledge of V, then the least square solution of regressing $V'Y \in \mathbb{R}^r$ on $X \in \mathbb{R}^p$ is

$$\begin{aligned}
U\Lambda &= \underset{L \in \mathbb{R}^{p \times r}}{\operatorname{argmin}} \, \mathbb{E}\|Y'V - X'L\|_{\mathrm{F}}^2 \\
&= \underset{L \in \mathbb{R}^{p \times r}}{\operatorname{argmin}} \, \mathsf{Tr}(L'\Sigma_x L) - 2\,\mathsf{Tr}(L'\Sigma_{xy}V) + \mathsf{Tr}(V'\Sigma_y V) \\
&= \underset{L \in \mathbb{R}^{p \times r}}{\operatorname{argmin}} \, \mathsf{Tr}(L'\Sigma_x L) - 2\,\mathsf{Tr}(L'\Sigma_{xy}V),
\end{aligned}\tag{14}$$

where the expectation is with respect to the distribution $(X', Y')' \sim N_{p+m}(0, \Sigma)$. The second equality results from taking expectation over each of the three terms in the expansion of the square Frobenius norm, and the last equality holds since $\mathsf{Tr}(V'\Sigma_y V)$ does not involve the argument to be optimized over. Comparing (10) with (14), it is clear that (10) is a sparsity constrained version of (14) where the knowledge of V and the covariance matrix Σ are replaced by the initial estimator $\widehat{V}^{(0)}$ and sample covariance matrix from an independent sample. Therefore, $\widehat{U}^{(1)}$ can be viewed as an estimator of $U\Lambda$. Hence, a final normalization step is taking in (11) to transform it to an estimator of U.

Under assumption (12), Theorem 3.2 shows that it is possible to achieve a high probability bound for prediction loss in U that does not depend on any parameter related to V. The optimality of this upper bound can be justified by the following minimax lower bound.

Theorem 3.3. *Assume that* $r \leq \frac{s_u \wedge s_v}{2}$. *Then there exists some constant* $C > 0$ *only depending on* M *and an absolute constant* $c_0 > 0$, *such that*

$$\inf_{\widehat{U}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\left(L(\widehat{U}, U) \geq c_0 \wedge \frac{C}{n\lambda^2} s_u \left(r + \log \frac{ep}{s_u} \right) \right) \geq 0.8,$$

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

By Theorem 3.2 and Theorem 3.3, we conclude that (13) is the minimax rate of the problem whenever it is upper bounded by a constant.

4 Adaptive and Computationally Efficient Estimation

The study in Section 3 determines the minimax rates for estimating U under the prediction loss. However, there are two drawbacks of the procedure (9) – (11). One is that the procedure requires the knowledge of the sparsity levels s_u and s_v. It is thus not adaptive. The other is that in both stages one needs to conduct exhaustive search over all subsets of given sizes in the optimization problems (9) and (10), and hence the computation cost is formidable.

In this section, we overcome both drawbacks by proposing a convex programming approach towards sparse CCA. The procedure is named as CoLaR, standing for Convex programming with group-Lasso Refinement. It is not only computationally feasible but also achieves the minimax estimation error rates adaptively over a large collection of parameter spaces under an additional sample size condition. In what follows, we introduce the procedure in Section 4.1 and then present its theoretical guarantee in Section 4.2. The issues related to the additional sample size condition will be discussed in more detail in the subsequent Section 5.

4.1 Estimation scheme

The basic principle underlying the computationally feasible estimation scheme is to seek tight convex relaxations of the combinatorial programs (9) – (10). In what follows, we introduce convex relaxations for the two stages in order. As in Section 3, we assume that the data is split into three batches $\mathcal{D}_0, \mathcal{D}_1$ and $\mathcal{D}_2$ of equal sizes and for $j = 0, 1, 2$, let $\widehat{\Sigma}_x^{(j)}, \widehat{\Sigma}_y^{(j)}$ and $\widehat{\Sigma}_{xy}^{(j)}$ be defined as before.

First stage By the definition of trace inner product, the objective function in (9) can be rewritten as $\mathsf{Tr}(L'\widehat{\Sigma}_{xy}R) = \langle \widehat{\Sigma}_{xy}, LR' \rangle$. Since it is linear in $F = LR'$, this suggests treating LR' as a single argument rather than optimizing over L and R separately. Next, the support size constraints $|\mathrm{supp}(L)| \leq s_u, |\mathrm{supp}(R)| \leq s_v$ imply that the vector ℓ_0 norm $\|LR'\|_0 \leq s_u s_v$. Applying the convex relaxation of ℓ_0 norm by ℓ_1 norm and including it as a Lagrangian term, we are led to consider a new objective function

$$\max_{F \in \mathbb{R}^{p \times m}} \langle \widehat{\Sigma}_{xy}^{(0)}, F \rangle - \rho \|F\|_1, \tag{15}$$

where F serves as a surrogate for LR', $\|F\|_1 = \sum_{i \in [p], j \in [m]} |F_{ij}|$ denotes the vector ℓ_1 norm of the matrix argument, and ρ is a penalty parameter controlling sparsity. Note that (15) is the maximization problem of a concave function, which becomes a convex program if the constraint set is convex. Under the identity $F = LR'$, the normalization constraint in (9) reduces to

$$(\widehat{\Sigma}_x^{(0)})^{1/2} F (\widehat{\Sigma}_y^{(0)})^{1/2} \in \mathcal{O}_r = \{AB' : A \in O(p, r), B \in O(m, r)\}. \tag{16}$$

Naturally, we relax it to $(\widehat{\Sigma}_x^{(0)})^{1/2} F (\widehat{\Sigma}_y^{(0)})^{1/2} \in \mathcal{C}_r$ where

$$\mathcal{C}_r = \{G \in \mathbb{R}^{p \times m} : \|G\|_* \leq r, \|G\|_{\mathrm{op}} \leq 1\} = \mathrm{conv}(\mathcal{O}_r) \tag{17}$$

is the smallest convex set containing $\mathcal{O}_r$. For a proof of (17), see Section 9.4 in the supplementary material. Combining (15) – (17), we use the following convex program for the first stage in our adaptive estimation scheme:

$$\max_{F \in \mathbb{R}^{p \times m}} \quad \langle \widehat{\Sigma}_{xy}^{(0)}, F \rangle - \rho \|F\|_1$$
$$\text{subject to} \quad \|(\widehat{\Sigma}_x^{(0)})^{1/2} F (\widehat{\Sigma}_y^{(0)})^{1/2}\|_* \leq r, \ \|(\widehat{\Sigma}_x^{(0)})^{1/2} F (\widehat{\Sigma}_y^{(0)})^{1/2}\|_{\mathrm{op}} \leq 1. \tag{18}$$

This optimization problem can be solved by the Alternating Direction Method with Multipliers (ADMM) [14, 10]. For details, see Section 10 in the supplementary material.

Remark 4.1. A related but different convex relaxation was proposed in [31] for the sparse PCA problem, where the set of all rank r projection matrices (which are symmetric) is relaxed to its convex hull – the Fantope $\{P : \mathrm{Tr}(P) = r, 0 \preceq P \preceq I_p\}$. Such an idea is not directly applicable in the current setting due to the asymmetric nature of the matrices included in the set $\mathcal{O}_r$ in (16).

Remark 4.2. The risk of the solution to (18) for estimating UV', as we shall see in Theorem 4.1 below, is sub-optimal compared to the optimal rates determined in [16] and Theorem 3.3. Nonetheless, it leads to a reasonable estimator for the subspaces spanned by first r left and right canonical directions under a sample size condition, which is sufficient for the purpose of achieving the optimal estimation rates for U and V in the second stage refinement to be introduced below. In some sense, a further refinement of the optimizer in (18) is indispensable for achieving optimal statistical performance. Actually, the improvement by the second stage can be considerable as to be revealed by both Theorems 4.1 and 4.2 below and the simulation results reported in Section 11 in the supplementary material.

Second stage Now we turn to convex relaxation to (10) in the second stage. By the discussion following Theorem 3.2, if we view the rows of L as groups, then (10) becomes a least square problem with a constrained number of active groups. A well-known convex relaxation for such problems is the group-Lasso [40] where the number of active groups constraint is relaxed by bounding the sum of ℓ_2 norms of the coefficient vector of each group. Let $\widehat{A}$ be the solution to (18) and $\widehat{U}^{(0)}$ (resp. $\widehat{V}^{(0)}$) be the matrix consisting of its first r left (resp. right) singular vectors. Thus, in the second stage of the adaptive estimation scheme, we propose to solve the following group-Lasso problem:

$$\min_{L \in \mathbb{R}^{p \times m}} \mathrm{Tr}(L' \widehat{\Sigma}_x^{(1)} L) - 2 \, \mathrm{Tr}(L' \widehat{\Sigma}_{xy}^{(1)} \widehat{V}^{(0)}) + \rho_u \sum_{j=1}^{p} \|L_{j\cdot}\|, \tag{19}$$

where $\sum_{j=1}^{p} \|L_{j\cdot}\|$ is the group sparsity penalty, defined as the sum of the ℓ_2 norms of all the row vectors in L, and ρ_u is a penalty parameter controlling sparsity. Note that the group

sparsity penalty is crucial, since if one uses an ℓ_1 penalty instead, only a sub-optimal rate can be achieved. Suppose the solution to (19) is $\widehat{U}^{(1)}$, then our final estimator in the adaptive estimation scheme is its normalized version

$$\widehat{U} = \widehat{U}^{(1)}((\widehat{U}^{(1)})'\widehat{\Sigma}_x^{(2)}\widehat{U}^{(1)})^{-1/2}. \tag{20}$$

As in Section 3, the reason of using sample splitting in the estimation scheme is only for the technical arguments in the proof. Simulation results in Section 11 in the supplementary material show that using the whole dataset repeatedly in (18) – (20) yields satisfactory performance.

4.2 Theoretical guarantees

We first state the upper bound for the solution $\widehat{A}$ to the convex program (18).

Theorem 4.1. *Assume that*

$$n \geq C_1 \frac{s_u s_v \log(p+m)}{\lambda^2}, \tag{21}$$

for some sufficiently large constant $C_1 > 0$. For any constant $C' > 0$, there exist positive constants γ_1, γ_2 and C only depending on M and C', such that when $\rho = \gamma\sqrt{\frac{\log(p+m)}{n}}$ for $\gamma \in [\gamma_1, \gamma_2]$,

$$\|\widehat{A} - UV'\|_{\mathrm{F}}^2 \leq C s_u s_v \rho^2/\lambda^2,$$

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

Note that the error bound in Theorem 4.1 can be much larger than the optimal rate for joint estimation of UV' established in Theorem 3.1 and [16]. Nonetheless, under the sample size condition (21), it still ensures that $\widehat{A}$ is close to UV' in Frobenius norm distance. This fact, together with the proposed refinement scheme (19) – (20), guarantees the optimal rates of convergence for the estimator (20) as stated in the following theorem.

Theorem 4.2. *Assume (21) holds for some sufficiently large $C_1 \geq 0$. For any $C' > 0$, there exist constants γ and γ_u only depending on C', C_1 and M such that if we set $\rho = \gamma'\sqrt{\frac{\log(p+m)}{n}}$ and $\rho_u = \gamma'_u\sqrt{\frac{r+\log p}{n}}$ for any $\gamma' \in [\gamma, C_2\gamma]$ and $\gamma'_u \in [\gamma_u, C_2\gamma_u]$ for some absolute constant $C_2 > 0$, there exists a constant $C > 0$ only depending on C', C_1, C_2 and M, such that*

$$L(\widehat{U}, U) \leq C\frac{s_u(r + \log p)}{n\lambda^2},$$

with $\mathbb{P}$-probability at least $1 - \exp(-C'(s_u + \log(ep/s_u))) - \exp(-C'(s_v + \log(em/s_v))) - \exp(-C'(r + \log(p \wedge m)))$ uniformly over $\mathbb{P} \in \mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M)$.

By Theorem 3.3, the rate in Theorem 4.2 is optimal. By Theorem 4.1 and Theorem 4.2, the choices of the penalty parameters ρ and ρ_u in (18) and (19) do not depend on s_u or s_v. Therefore, the proposed estimation scheme (18) – (20) achieves the optimal rate adaptively over sparsity levels.

We conclude this section with two important remarks.

12

Remark 4.3. The group sparsity penalty used in the second stage (19) plays an important role in achieving the optimal rate $\frac{s_u(r+\log p)}{n\lambda^2}$. If we simply use an ℓ_1 penalty, then we will obtain the rate $\frac{rs_u\log p}{n\lambda^2}$, which is clearly sub-optimal.

Remark 4.4. Comparing Theorem 3.2 with Theorem 4.2, the adaptive estimation scheme achieves the optimal rates of convergence for a smaller collection of parameter spaces of interest due to the more restrictive sample size condition (21). This inevitably invites the question: Is such a condition necessary? In Section 5, we show evidence that a condition of this kind is unavoidable for any polynomial time algorithm to produce a consistent estimator even for Gaussian data based on the conjectured hardness of the Planted Clique detection problem. Therefore, the sample size condition is in some sense necessary for any computationally tractable algorithm.

5   Computational Lower Bounds

In this section, we provide evidence that the sample size condition (21) imposed on the adaptive estimation scheme in Theorems 4.1 and 4.2 is essentially unavoidable for any computationally feasible estimator. To be specific, we show that for a sequence of parameter spaces in (5) – (6), if the condition is violated, then any computationally efficient estimator of sparse canonical directions leads to an efficient algorithm for the Planted Clique detection problem in a regime where it is believed to be computationally intractable.

Let N be a positive integer and $k \in [N]$. We denote by $\mathcal{G}(N, 1/2)$ the Erdős-Rényi graph on N vertices where each edge is drawn independently with probability $1/2$, and by $\mathcal{G}(N, 1/2, k)$ the random graph generated by first sampling from $\mathcal{G}(N, 1/2)$ and then selecting k vertices uniformly at random and forming a clique of size k on these vertices. For an adjacency matrix $A \in \{0, 1\}^{N \times N}$ of an instance from either $\mathcal{G}(N, 1/2)$ or $\mathcal{G}(N, 1/2, k)$, the Planted Clique detection problem of parameter (N, k) refers to testing the following hypotheses

$$H_0^G : A \sim \mathcal{G}(N, 1/2) \quad \text{v.s.} \quad H_1^G : A \sim \mathcal{G}(N, 1/2, k). \tag{22}$$

To form the connection of the Planted Clique problem to sparse CCA, let us first define a sparse canonical correlation detection problem. Denoting the distribution $N_{p+m}(0, I_{p+m})$ by $\mathbb{P}_0$, we consider the following detection problem:

$$\begin{aligned} H_0^C &: \{(X_i', Y_i')'\}_{i=1}^n \sim \mathbb{P}_0^n \quad \text{v.s.} \\ H_1^C &: \{(X_i', Y_i')'\}_{i=1}^n \sim \mathbb{P} \in \mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M), \end{aligned} \tag{23}$$

where the parameter space in H_1^C is defined as in (6). To establish the connection between (22) and (23), we shall propose a four-step reduction scheme from (22) to (23) where $n \asymp N$, $s_u = s_v \asymp k$, $r = 1$, $\lambda \asymp k^2/N$ up to a sub-polynomial factor and $p = m$ with $\log p \asymp \log n$. From a given adjacency matrix A, we are able to generate observations $\{(X_i', Y_i')'\}_{i=1}^n$, such that when A follows $\mathcal{G}(N, 1/2)$, the distribution of $\{(X_i', Y_i')'\}_{i=1}^n$ is close to $\mathbb{P}_0^n$ in total variation, and when A follows $\mathcal{G}(N, 1/2, k)$, the distribution of $\{(X_i', Y_i')'\}_{i=1}^n$ is close in total

variation to a mixture of distributions in $\mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M)$. When there is a good estimator of the leading canonical direction, we are able to test between H_0^C and H_1^C, which immediately leads to a good test between H_0^G and H_1^G. Note that the reduction scheme proposed in this paper to connect (22) and (23) is directly targeted for the Gaussian sparse CCA model, and does not require any parameter space enlargement as was done in [6, 33] for the related sparse PCA problem. A delicate procedure is incorporated in the proposed reduction scheme to generate nearly Gaussian distributed i.i.d. random vectors $\{(X_i', Y_i')'\}_{i=1}^n$ from a Bernoulli random matrix A.

It is widely believed that when $k = o(\sqrt{N})$, the Planted Clique detection problem (22) cannot be solved by any randomized polynomial-time algorithm. According to the aforementioned correspondence between (N, k) and $(n, s_u, s_v, p, m, r, \lambda)$ by our reduction scheme, the hard regime $k = o(\sqrt{N})$ for the Planted Clique problem corresponds to the regime of sparse CCA problem where the sample size condition (21) is violated. Hence, the computational hardness of the Planted Clique problem implies the computational hardness of sparse CCA. The hypothesized hardness of Planted Clique problem can be formalized into the following hypothesis.

Hypothesis A. *For any sequence* $k = k(N)$ *such that* $\limsup_{N \to \infty} \frac{\log k}{\log N} < \frac{1}{2}$ *and any randomized polynomial-time test* ψ,

$$\liminf_{N \to \infty} \left(\mathbb{P}_{H_0^G} \psi + \mathbb{P}_{H_1^G} (1 - \psi) \right) \geq \frac{2}{3}.$$

Evidences supporting this hypothesis have been provided in [29, 15]. Recently, computational lower bounds in several statistical problems have been established by assuming the above hypothesis and its close variants, including sparse PCA detection [6] and estimation [33], submatrix detection [26] and community detection [18].

In what follows, Section 5.1 introduces the asymptotically equivalent discretized model and states the rigorous computational lower bounds for the discretized sparse CCA problem. The key step in establishing the lower bound is a randomized polynomial time reduction which maps any solution to the sparse CCA estimation problem to a solution to the Planted Clique detection problem, where dealing with discrete data is necessary for rigorous complexity theoretic investigation under Turing machine models [2]. To convey the main ideas in a more transparent way, we present a sketch of the reduction scheme in Section 5.2 ignoring the discretization issue. A rigorous treatment is deferred to Section 8.3 in the supplement. A slight variant of the proposed reduction scheme leads to a computational lower bound for sparse PCA under the Gaussian spiked covariance model. For details, see Section 7 in the supplement.

5.1 Asymptotically equivalent discretization and hardness of sparse CCA

To formally address the computational complexity issue in a continuous statistical model, we adopt the framework of asymptotically equivalent discretization proposed in [26]. The asymptotically equivalent discretized model allows computational complexity to be well-defined,

while preserving the statistical difficulty of the original continuous problem. For any $t \in \mathbb{N}$, define the function $[\cdot]_t : \mathbb{R} \to 2^{-t}\mathbb{Z}$ by

$$[x]_t = 2^{-t}\lfloor 2^t x \rfloor. \tag{24}$$

For any vector $v = (v_i)$ and any matrix $R = (R_{ij})$, $[v]_t = ([v_i]_t)$ and $[R]_t = ([R_{ij}]_t)$. Let $\mathcal{E}_M^{(p,n)} = \{\mathcal{L}(X_1, \ldots, X_n) : X_i \overset{iid}{\sim} N_p(\mu, \Sigma), M^{-1} \le \sigma_{\min}(\Sigma) \le \sigma_{\max}(\Sigma) \le M\}$, and $\mathcal{E}_M^{(p,n,t)} = \{\mathcal{L}([X_1]_t, \ldots, [X_n]_t) : \mathcal{L}(X_1, \ldots, X_n) \in \mathcal{E}_M^{(p,n)}\}$. The following lemma bounds the Le Cam distance [23] and hence establishes the asymptotic equivalence of multivariate Gaussian distribution and its appropriate discretization.

Lemma 5.1. *When* $2^t t^{-1/2} \ge 2(pM)^{3/2}$, *the Le Cam distance* $\Delta(\mathcal{E}_M^{(p,n)}, \mathcal{E}_M^{(p,n,t)}) \le n(pM)^{3/2} t^{1/2} 2^{-t}$.

Now define the discretized sparse CCA probability space: for any $t \in \mathbb{N}$,

$$\mathcal{P}^t(n, s_u, s_v, p, m, r, \lambda; M) = \{\mathcal{L}([X]_t, [Y]_t) : \mathcal{L}(X, Y) \in \mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M)\}.$$

The following theorem gives the computational lower bound for the sparse CCA estimation problem considered in the present paper.

Theorem 5.1. *Suppose that Hypothesis A holds and that as* $n \to \infty$, $p = m$ *satisfying* $2n \le p \le n^a$ *for some constant* $a > 1$, $s_u = s_v$, $n(\log n)^5 \le cs_u^4$ *for some sufficiently small* $c > 0$, *and* $\lambda = \frac{s_u s_v}{9720 n (\log(12n))^2}$. *If for some* $\delta \in (0, 1)$,

$$\liminf_{n \to \infty} \frac{(s_u s_v)^{1-\delta} \log(p+m)}{n\lambda^2} > 0, \tag{25}$$

then for any randomized polynomial-time estimator $\widehat{u}$,

$$\liminf_{n \to \infty} \sup_{\mathbb{P} \in \mathcal{P}^t(n, s_u, s_v, p, m, 1, \lambda; 4)} \mathbb{P}\left\{L(\widehat{u}, u) > \frac{1}{3 \times 32^2}\right\} > \frac{1}{4}, \tag{26}$$

where $t = \lceil 4 \log_2(p + m + n) \rceil$.

Let us abbreviate $\mathcal{P}^t(n, s_u, s_v, p, m, 1, \lambda; 4)$ and $\mathcal{P}(n, s_u, s_v, p, m, 1, \lambda; 4)$ by $\mathcal{P}^t$ and $\mathcal{P}$. On one hand, when $t = \lceil 4 \log_2(p + m + n) \rceil$, Lemma 5.1 implies that $\mathcal{P}^t$ is asymptotically equivalent to $\mathcal{P}$. This is because $\mathcal{P}^t \subset \mathcal{E}_5^{(p+m,n,t)}$ and $\mathcal{P} \subset \mathcal{E}_5^{(p+m,n)}$, which implies $\Delta(\mathcal{P}^t, \mathcal{P}) \le \Delta(\mathcal{E}_5^{(p+m,n,t)}, \mathcal{E}_5^{(p+m,n)}) \to 0$ with the given t. In addition, following the lines of the proofs of Theorems 3.1–3.2 and 4.1–4.2, one can show that the same upper bounds continue to hold when we apply the estimator (9)–(11) and the adaptive procedure (18)–(20) on the discrete data directly. In summary, results in Sections 3 and 4 continue to hold for appropriately discretized problems. On the other hand, Theorem 5.1 provides a sequence of asymptotically equivalent discretized models under which the condition (21) is not only sufficient but also *necessary* (up to a sub-polynomial factor) for any randomized polynomial-time estimator to be consistent. Therefore, the computationally feasible adaptive estimation scheme in Section 4 does not require excessively strong condition to achieve optimal rates of convergence.

15

5.2   A sketch of the reduction scheme

The key step in establishing the computational lower bounds in Theorem 5.1 is a randomized polynomial-time reduction that maps any solution to the sparse CCA estimation problem to a solution to the Planted Clique detection problem. To better explain the main ideas, we present below the construction for the continuous case. A discretized reduction scheme is introduced in Section 8.3 in the supplement.

Preliminaries   We start with some notation. Consider integers k and N. Define

$$\delta_N = \frac{k}{N}, \quad \eta_N = \frac{k}{45 N (\log N)^2}. \tag{27}$$

For any $\mu \in \mathbb{R}$, let ϕ_μ denote the density function of the $N(\mu, 1)$ distribution, and define

$$\bar{\phi}_\mu = \frac{1}{2}(\phi_\mu + \phi_{-\mu}). \tag{28}$$

Next, let $\widetilde{\Phi}_0$ be the restriction of the $N(0,1)$ distribution on the interval $[-3\sqrt{\log N}, 3\sqrt{\log N}]$. For any $|\mu| \leq 3\sqrt{\eta_N \log N}$, define two probability distributions $\mathcal{F}_{\mu,0}$ and $\mathcal{F}_{\mu,1}$ with densities

$$f_{\mu,0}(x) \;=\; M_0 \left(\phi_0(x) - \delta_N^{-1} [\bar{\phi}_\mu(x) - \phi_0(x)] \right) \mathbf{1}_{\{|x| \leq 3\sqrt{\log N}\}}, \tag{29}$$

$$f_{\mu,1}(x) \;=\; M_1 \left(\phi_0(x) + \delta_N^{-1} [\bar{\phi}_\mu(x) - \phi_0(x)] \right) \mathbf{1}_{\{|x| \leq 3\sqrt{\log N}\}}, \tag{30}$$

where the M_i's are normalizing constants such that $\int_{\mathbb{R}} f_{\mu,i} = 1$ for $i = 0, 1$. It can be verified that $f_{\mu,i}$ are properly defined probability density function when $|\mu| \leq 3\sqrt{\eta_N \log N}$. For details, see Section 8.1. The reason for introducing the above two distributions is to match specific mixtures of them to ϕ_0 and $\bar{\phi}_\mu$ respectively as summarized in the following lemma.

Lemma 5.2. *There exists an absolute constant $C > 0$, such that for all integers $N \geq 12$, $k \leq N/12$ and all $|\mu| \leq 3\sqrt{\eta_N \log N}$,*

$$\mathsf{TV}(h_{\mu,0}, \phi_0) \leq C N^{-3} \quad and \quad \mathsf{TV}(h_{\mu,1}, \bar{\phi}_\mu) \leq C N^{-3},$$

where $h_{\mu,0} = \frac{1}{2}(f_{\mu,0} + f_{\mu,1})$ and $h_{\mu,1} = \delta_N f_{\mu,1} + (1 - \delta_N)\frac{1}{2}(f_{\mu,0} + f_{\mu,1})$.

Reduction   We now propose our approach to turning an estimator for the sparse CCA problem to a testing procedure for (22).

Let $A \in \{0,1\}^{N \times N}$ be an adjacency matrix sampled either from H_0^G or H_1^G. Let $n \leq N/12$, and we first construct n pairs of random vectors (X_i, Y_i) where $X_i \in \mathbb{R}^p, Y_i \in \mathbb{R}^m$ with $p = m \geq 2n$. The goal here is to ensure that the joint distribution of $\{(X_i, Y_i)\}_{i=1}^n$ is close to H_0^C in (23) when $A \sim H_0^G$ and to a mixture of the distributions in H_1^C when $A \sim H_1^G$. This would allow us to turn any test for (23) to a test for (22), and the remaining job would be to turn an estimator of the canonical directions to a test for the sparse CCA hypotheses (23). To this end, we construct n auxiliary random vectors W_i, which are asymptotically independent

16

of $\{(X_i, Y_i)\}_{i=1}^n$. When $A \sim H_0^G$, the W_i's are close in distribution to i.i.d. $N_p(0, I_p)$ vectors. When $A \sim H_1^G$, the leading eigenvector of the covariance matrix of W_i is essentially identical to the leading canonical directions between the (X_i, Y_i)'s. With the aid of the W_i's, we are able to turn any sparse CCA estimator to a test for the sparse CCA hypothesis (23).

More precisely, the reduction scheme consists of the following four steps in order. The first three steps generate $\{(X_i, Y_i)\}_{i=1}^n$ and $\{W_i\}_{i=1}^n$ and the last step constructs the test.

1. **Initialization**. Generate i.i.d. random variables $\xi_1, \ldots, \xi_{2n} \sim \widetilde{\Phi}_0$. Set

$$\mu_i = \eta_N^{1/2} \xi_i, \quad i = 1, \ldots, 2n. \tag{31}$$

2. **Gaussianization**. Generate two matrices $B_0, B_1 \in \mathbb{R}^{2n \times 2n}$ where conditioning on the μ_i's, all the entries are mutually independent satisfying

$$\mathcal{L}((B_0)_{ij}|\mu_i) = \mathcal{F}_{\mu_i,0} \quad \text{and} \quad \mathcal{L}((B_1)_{ij}|\mu_i) = \mathcal{F}_{\mu_i,1}. \tag{32}$$

Let $A_0 \in \{0,1\}^{2n \times 2n}$ be the lower–left $2n \times 2n$ submatrix of the matrix A. Generate a matrix $W \in \mathbb{R}^{2n \times p}$ where for each $i \in [2n]$, if $j \in [2n]$, then we set

$$W_{ij} = (B_0)_{ij} (1 - (A_0)_{ij}) + (B_1)_{ij} (A_0)_{ij}. \tag{33}$$

If $2n < j \leq p$, we let W_{ij} be an independent draw from $N(0,1)$.

3. **Sample Generation**. For $i \in [2n]$, let $W_i = (W_{i1}, \ldots, W_{ip})'$ be i-th row vector of W. Then for $i = 1, \ldots, n$, we generate independent standard normal vector $Z_i \sim N_p(0, I_p)$. Define

$$X_i = \frac{1}{\sqrt{2}}(W_{n+i} + Z_i), \qquad Y_i = \frac{1}{\sqrt{2}}(W_{n+i} - Z_i), \tag{34}$$

and let $X = [X_1', \ldots, X_n']' \in \mathbb{R}^{n \times p}$ and $Y = [Y_1', \ldots, Y_n']' \in \mathbb{R}^{n \times m}$.

4. **Test Construction**. Let $\widehat{u} = \widehat{u}(X, Y)$ be the estimator of the first canonical correlation direction by treating $\{(X_i, Y_i)\}_{i=1}^n$ as data. We reject H_0^G if

$$\frac{\widehat{u}'(\frac{1}{n} \sum_{i=1}^n W_i W_i')\widehat{u}}{\|\widehat{u}\|^2} \geq 1 + \frac{1}{4} k \, \eta_N. \tag{35}$$

We now discuss in more detail how the reduction scheme achieves its goal. For simplicity, focus on the case where $p = m = 2n$. Let $\epsilon = (\epsilon_1, ..., \epsilon_{2n})$ be a binary vector where ϵ_i is the indicator of whether the i-th row of A_0 belongs to the planted clique or not, and $\gamma = (\gamma_1, ..., \gamma_{2n})$ the indicators of the columns of A_0. In what follows, we discuss the distributions of W, X and Y when $A \sim H_0^G$ and H_1^G, respectively.

When $A \sim H_0^G$, the ϵ_i's and γ_j's are all zeros. In this case, we can verify that the entries of W are mutually independent and for each (i, j) the marginal distribution of W_{ij} is close to the $N(0,1)$ distribution by Lemma 5.2. Hence, the rows of W are close to i.i.d. random

vectors from the $N_p(0, I_p)$ distribution. This, together with (34), further implies that the X_i's and Y_i's are close to i.i.d. random vectors from the $N_p(0, I_p)$ distribution, and they are independent of $W_1, \ldots, W_n$. Since $\widehat{u}$ is independent of $\{W_i\}_{i=1}^n$, the LHS of (35) is close in distribution to a χ_n^2 random variable scaled by n which concentrates around its expected value one. Indeed, the LHS is upper bounded by $1 + O(\sqrt{\log(n)/n})$ with high probability.

If $A \sim H_1^G$, then the (i, j)-th entry of A_0 is an edge in the planted clique if and only if $\epsilon_i = \gamma_j = 1$. Moreover, the joint distribution of $\{\epsilon_1, \ldots, \epsilon_{2n}, \gamma_1, \ldots, \gamma_{2n}\}$ is close to that of $4n$ i.i.d. Bernoulli random variables $\{\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_{2n}, \widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_{2n}\}$ with success probability $\delta_N = k/N$. To get the intuition, suppose that the indicators of whether the rows and the columns belong to the planted clique are i.i.d. Bernoulli(δ_N) variables $\{\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_{2n}, \widetilde{\gamma}_1, \ldots, \widetilde{\gamma}_{2n}\}$. Then, Lemma 5.2 implies that on one hand, conditioning on $\widetilde{\gamma}_j = 0$, for any $i \in [2n]$, the conditional distribution of $(W_{ij}|\widetilde{\gamma}_j = 0)$, after integrating over the conditional distribution of $\widetilde{\epsilon}_i$, μ_i and $(A_0)_{ij}$, is close in total variation to the $N(0, 1)$ distribution. On the other hand, conditioning on $\widetilde{\gamma}_j = 1$, for any $i \in [2n]$, the conditional distribution of $(W_{ij}|\widetilde{\gamma}_j = 1)$ is close in total variation to $N(0, 1 + \eta_N)$. Therefore, conditioning on $\widetilde{\gamma}$ the distribution of the W_i's is close to that of $2n$ i.i.d. random vectors sampled from the distribution

$$N_p(0, \tau\theta\theta' + I_p), \quad \text{where } \theta = \widetilde{\gamma}/\|\widetilde{\gamma}\| \text{ and } \tau = \eta_N\|\widetilde{\gamma}\|^2, \tag{36}$$

which is of the form of a Gaussian spiked covariance model used in the sparse PCA literature. Here, the leading eigenvector θ has sparsity level $|\text{supp}(\theta)| = |\text{supp}(\widetilde{\gamma})| = \sum_j \widetilde{\gamma}_j$, which concentrates around its mean value $n\delta_N \asymp k$ if $N \asymp n$. In addition, the sample generation (34) ensures that the $(X_i', Y_i')' \in \mathbb{R}^{p+m}$ are close to i.i.d. random vectors sampled from $N_{p+m}(0, \Sigma)$ where

$$\Sigma_x = \Sigma_y = \frac{\tau}{2}\theta\theta' + I_p, \ \Sigma_{xy} = \Sigma_x(\lambda uv')\Sigma_y \tag{37}$$

with $u = v = \frac{\theta}{\sqrt{\tau/2+1}}$, $\lambda = \frac{\tau/2}{\tau/2+1}$. This is a special case of the Gaussian canonical pair model (2). In addition, the (X_i, Y_i) pairs are (asymptotically) independent of $W_1, \ldots, W_n$. Thus, if $\widehat{u}$ estimates u in (37) well, then $\widehat{u}/\|\widehat{u}\|$ is close to θ (up to a sign change), the leading eigenvector of the covariance matrix of $W_1, \ldots, W_n$. Thus, the LHS of (35) should exceed $1 + O(\sqrt{\log(n)/n})$ under the alternative hypothesis and hence yield a test with small error for the Planted Clique problem (22).

The materialization of the foregoing discussion leads to the following result which demonstrates quantitatively that a decent estimator of the leading canonical correlation direction results in a good test (by applying the reduction (31) – (35)) for the Planted Clique detection problem (22).

Theorem 5.2. *For some sufficiently small constant $c > 0$, assume $\frac{k^2}{N(\log N)^2} \vee \frac{N(\log N)^5}{k^4} \leq c$, $cN \leq n \leq N/12$ and $p \geq 2n$. Then, for any $\widehat{u}$ such that*

$$\sup_{\mathbb{P} \in \mathcal{P}(n, 3k/2, 3k/2, p, p, 1, k\eta_N/8; 4)} \mathbb{P}\left\{L(\widehat{u}, u) > \frac{1}{3 \times 32^2}\right\} \leq \beta, \tag{38}$$

18

the test ψ defined by (31) – (35) *satisfies*

$$\mathbb{P}_{H_0^G}\psi + \mathbb{P}_{H_1^G}(1 - \psi) < \beta + \frac{4n}{N} + C(n^{-1} + N^{-1} + e^{-C'k}),$$

for sufficiently large n with some constants $C, C' > 0$.

Remark 5.1. To bridge the gap between Theorem 5.2 and the desired result in Theorem 5.1, we need to turn the above sketch (31) - (34) into a randomized polynomial-time reduction for discrete data. To this end, the major modification is to replace the random number generation in steps 1–3 with their discrete counterparts. For details, see Section 8.3.2 in the supplement.

Remark 5.2. This section mainly studies computational lower bounds for sparse CCA estimation. Similar results also hold for sparse CCA detection. Indeed, if we have a testing procedure for the sparse CCA hypotheses in (23), we can replace the fourth reduction step with directly applying the test to the (X_i, Y_i) vectors constructed in (33). A simple modification of the proof of Theorem 5.1 then leads to the proof of computational lower bounds for sparse CCA detection.

Remark 5.3. As we have hinted in (36), a slight variant of the reduction leads to a computational lower bound for sparse PCA under the Gaussian spiked covariance model. This allows us to close the gap in sparse PCA computational lower bounds left by [6] and [33]. A detailed discussion is given in Section 7 in the supplement.

6 Proofs

This section presents proofs of Theorems 4.1 and 4.2. The proofs of the other theoretical results are given in the supplement.

6.1 Proof of Theorem 4.1

Before presenting the proof, we state some technical lemmas. The proofs of all the lemmas are given in Section 9.4 in the supplement. First, note that the estimator is normalized with respect to $\widehat{\Sigma}_x^{(0)}$ and $\widehat{\Sigma}_y^{(0)}$, while the truth U and V is normalized with respect to Σ_x and Σ_y. To address this issue, we normalize the truth with respect to $\widehat{\Sigma}_x^{(0)}$ and $\widehat{\Sigma}_y^{(0)}$ to obtain $\widetilde{U} = U(U'\widehat{\Sigma}_x^{(0)}U)^{-1/2}$ and $\widetilde{V} = V(V'\widehat{\Sigma}_y^{(0)}V)^{-1/2}$. Also define $\widetilde{\Lambda} = (U'\widehat{\Sigma}_x^{(0)}U)^{1/2}\Lambda(V'\widehat{\Sigma}_y^{(0)}V)^{1/2}$. For notational convenience, define

$$\epsilon_{n,u} = \sqrt{\frac{1}{n}\left(s_u + \log\frac{ep}{s_u}\right)}, \quad \epsilon_{n,v} = \sqrt{\frac{1}{n}\left(s_v + \log\frac{em}{s_v}\right)}. \tag{39}$$

The following lemma bounds the normalization effect.

19

Lemma 6.1. *Assume $\epsilon_{n,u}^2 + \epsilon_{n,v}^2 \leq c$ for some small constant $c > 0$. Then, for any $C' > 0$, there exists $C > 0$ only depending on C' such that*

$$\|\Sigma_x^{1/2}(\widetilde{U} - U)\|_{\mathrm{op}} \leq C\epsilon_{n,u}, \quad \|\Sigma_y^{1/2}(\widetilde{V} - V)\|_{\mathrm{op}} \leq C\epsilon_{n,v},$$
$$\|\widetilde{\Lambda} - \Lambda\|_{\mathrm{op}} \leq C(\epsilon_{n,u} + \epsilon_{n,v}),$$

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

Using the definitions of $\widetilde{U}$ and $\widetilde{V}$, let us state the following lemma, which asserts that the matrix $\widetilde{A} = \widetilde{U}\widetilde{V}'$ is feasible to the optimization problem (18).

Lemma 6.2. *Define $\widetilde{A} = \widetilde{U}\widetilde{V}'$. When $\widetilde{A}$ exists, we have*

$$\|(\widehat{\Sigma}_x^{(0)})^{1/2}\widetilde{A}(\widehat{\Sigma}_y^{(0)})^{1/2}\|_* = r \quad and \quad \|(\widehat{\Sigma}_x^{(0)})^{1/2}\widetilde{A}(\widehat{\Sigma}_y^{(0)})^{1/2}\|_{\mathrm{op}} = 1.$$

As was argued in Section 4.1, the set $\mathcal{C}_r$ is the convex hull of $\mathcal{O}_r$. To show that the relaxation $\mathcal{C}_r$ preserves the curvature of the original constraint $\mathcal{O}_r$, we need the following curvature lemma.

Lemma 6.3. *Let $F \in O(p, r)$, $G \in O(m, r)$, $K \in \mathbb{R}^{r \times r}$ and $D = \mathrm{diag}(d_1, ..., d_r)$ with $d_1 \geq ... \geq d_r > 0$. If E satisfies $\|E\|_{\mathrm{op}} \leq 1$ and $\|E\|_* \leq r$, then*

$$\langle FKG', FG' - E \rangle \geq \frac{d_r}{2}\|FG' - E\|_{\mathrm{F}}^2 - \|K - D\|_{\mathrm{F}}\|FG' - E\|_{\mathrm{F}}. \tag{40}$$

Define

$$\widetilde{\Sigma}_{xy} = \widehat{\Sigma}_x^{(0)}U\Lambda V'\widehat{\Sigma}_y^{(0)}. \tag{41}$$

Lemma 6.4 is instrumental in determining the proper value of the tuning parameter required in the program (18).

Lemma 6.4. *Assume $r\sqrt{\frac{\log(p+m)}{n}} \leq C_1$ 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*

$$\|\widehat{\Sigma}_{xy}^{(0)} - \widetilde{\Sigma}_{xy}\|_\infty \leq C\sqrt{\frac{\log(p+m)}{n}},$$

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

We also need a lemma on restricted eigenvalue. For any p.s.d. matrix B, define

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

The following lemma is adapted from Lemma 12 in [16], and its proof is omitted.

Lemma 6.5. *Assume* $\frac{1}{n}\big((k_u \wedge p)\log(ep/(k_u \wedge p)) + (k_v \wedge m)\log(em/(k_v \wedge m))\big) \le C_1$ *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 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*

$$M^{-1} - C\delta_u(k_u) \le \phi_{\min}^{\widehat{\Sigma}_x^{(j)}}(k_u) \le \phi_{\max}^{\widehat{\Sigma}_x^{(j)}}(k_u) \le M + C\delta_u(k_u),$$

$$M^{-1} - C\delta_v(k_v) \le \phi_{\min}^{\widehat{\Sigma}_y^{(j)}}(k_v) \le \phi_{\max}^{\widehat{\Sigma}_y^{(j)}}(k_v) \le M + C\delta_v(k_v),$$

with probability at least $1 - \exp\big(-C'(k_u \wedge p)\log(ep/(k_u \wedge p))\big) - \exp\big(-C'(k_v \wedge m)\log(em/(k_v \wedge m))\big)$, *for* $j = 0, 1, 2$.

Finally, we need a result on subspace distance. Recall that for a matrix F, P_F denotes the projection matrix onto its column subspace.

Lemma 6.6. *For any matrix* $F \in O(d, r)$ *and any matrix* $G \in \mathbb{R}^{d \times r}$, *we have*

$$\inf_W \|F - GW\|_F^2 = \frac{1}{2}\|P_F - P_G\|_F^2.$$

If furthermore, $G \in O(d, r)$, *then we have*

$$\inf_{W \in O(r,r)} \|F - GW\|_F^2 = \frac{1}{2}\|P_F - P_G\|_F^2.$$

Proof of Theorem 4.1. In the rest of this proof, we denote $\widehat{\Sigma}_x^{(0)}$, $\widehat{\Sigma}_y^{(0)}$ and $\widehat{\Sigma}_{xy}^{(0)}$ by $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$ and $\widehat{\Sigma}_{xy}$ for notational convenience. We also let $\Delta = \widehat{A} - \widetilde{A}$. The proof consists of two steps. In the first step, we are going to derive an upper bound for $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F$. In the second step, we derive a generalized cone condition and use it to lower bound $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F$ by a constant multiple of $\|\Delta\|_F$ and hence the upper bound on $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F$ leads to an upper bound on $\|\Delta\|_F$.

Step 1. By Lemma 6.1, $\widetilde{U}$ and $\widetilde{V}$ are well-defined with high probability. Thus, $\widetilde{A}$ is well-defined with high probability, and we have

$$\|\Sigma_x^{1/2}(\widetilde{A} - UV')\Sigma_y^{1/2}\|_{\mathrm{op}} \le C(\epsilon_{n,u} + \epsilon_{n,v}). \tag{42}$$

with probability at least $1 - \exp\left(-C'(s_u + \log(ep/s_u))\right) - \exp\left(-C'(s_v + \log(em/s_v))\right)$. According to Lemma 6.2, $\widetilde{A}$ is feasible. Then, by the definition of $\widehat{A}$, we have

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

After rearrangement, we have

$$-\langle \widetilde{\Sigma}_{xy}, \Delta \rangle \le \rho\Big(\|\widetilde{A}\|_1 - \|\widetilde{A} + \Delta\|_1\Big) + \langle \widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}, \Delta \rangle, \tag{43}$$

where $\widetilde{\Sigma}_{xy}$ is defined in (41). For the first term on the right hand side of (43), we have

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

21

For the second term on the right hand side of (43), we have $\langle \widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}, \Delta \rangle \leq ||\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}||_\infty ||\Delta||_1$. Thus when

$$\rho \geq 2||\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}||_\infty, \tag{44}$$

we have

$$-\langle \widetilde{\Sigma}_{xy}, \Delta \rangle \leq \frac{3\rho}{/} 2||\Delta_{S_u S_v}||_1 - \frac{\rho}{2}||\Delta_{(S_u S_v)^c}||_1. \tag{45}$$

Using Lemma 6.3, we can lower bound the left hand side of (45) as

$$
\begin{aligned}
-\langle \widetilde{\Sigma}_{xy}, \Delta \rangle &= \langle \widehat{\Sigma}_x^{1/2} U \Lambda V' \widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2} \rangle \\
&= \langle \widehat{\Sigma}_x^{1/2} \widetilde{U} \widetilde{\Lambda} \widetilde{V}' \widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2} \rangle \\
&\geq \frac{1}{2}\lambda_r ||\widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2}||_F^2 - \delta ||\widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A}) \widehat{\Sigma}_y^{1/2}||_F,
\end{aligned}
\tag{46}
$$

where $\delta = ||\widetilde{\Lambda} - \Lambda||_F$. Combining (45) and (46), we have

$$\lambda_r ||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F^2 \leq 3\rho ||\Delta_{S_u S_v}||_1 - \rho ||\Delta_{(S_u S_v)^c}||_1 + 2\delta ||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F \tag{47}$$

$$\leq 3\rho ||\Delta_{S_u S_v}||_1 + 2\delta ||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F. \tag{48}$$

Solving the quadratic equation (48) by Lemma 2 of [11], we have

$$||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F^2 \leq 6\rho ||\Delta_{S_u S_v}||_1 / \lambda_r + 4\delta^2 / \lambda_r^2. \tag{49}$$

Combining (47) and (49), we have

$$
\begin{aligned}
0 &\leq 3\rho ||\Delta_{S_u S_v}||_1 - \rho ||\Delta_{(S_u S_v)^c}||_1 + \delta^2 / \lambda_r + \lambda_r ||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F^2 \\
&\leq 9\rho ||\Delta_{S_u S_v}||_1 - \rho ||\Delta_{(S_u S_v)^c}||_1 + 5\delta^2 / \lambda_r,
\end{aligned}
\tag{50}
$$

which gives rise to the generalized cone condition that we are going to use in Step 2. Finally, by the bound $||\Delta_{S_u S_v}||_1 \leq \sqrt{s_u s_v} \rho ||\Delta_{S_u S_v}||_F$ and (49), we have

$$||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F^2 \leq 6\sqrt{s_u s_v} \rho ||\Delta_{S_u S_v}||_F / \lambda_r + 4\delta^2 / \lambda_r^2, \tag{51}$$

which completes the first step.

Step 2. By (50), we have obtained the following condition

$$||\Delta_{(S_u S_v)^c}||_1 \leq 9||\Delta_{S_u S_v}||_1 + \frac{5\delta^2}{\rho \lambda_r}. \tag{52}$$

Due to the existence of the extra term $5\delta^2/(\rho \lambda_r)$ on the RHS, we call it a *generalized cone condition*. In this step, we are going to lower bound $||\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}||_F$ by $||\Delta||_F$ on the generalized cone. Motivated by the argument in [8], 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 $|J_K| \leq t$),

22

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,

$$\|\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} \geq \|\widehat{\Sigma}_x^{1/2} \Delta_{\widetilde{J}} \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} - \sum_{k=2}^{K} \|\widehat{\Sigma}_x^{1/2} \Delta_{J_k} \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$$

$$\geq \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u + t) \phi_{\min}^{\widehat{\Sigma}_y}(s_v + t)} \|\Delta_{\widetilde{J}}\|_{\mathrm{F}} - \sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t) \phi_{\max}^{\widehat{\Sigma}_y}(t)} \sum_{k=2}^{K} \|\Delta_{J_k}\|_{\mathrm{F}}.$$

By the construction of J_k, we have

$$
\begin{aligned}
\sum_{k=2}^{K} \|\Delta_{J_k}\|_{\mathrm{F}} &\leq \sqrt{t} \sum_{k=2}^{K} \|\Delta_{J_k}\|_{\infty} \leq t^{-1/2} \sum_{k=2}^{K} \|\Delta_{J_{k-1}}\|_1 \leq t^{-1/2} \|\Delta_{(S_u S_v)^c}\|_1 \\
&\leq t^{-1/2} \left(9\|\Delta_{S_u S_v}\|_1 + \frac{5\delta^2}{\rho\lambda_r} \right) \leq 9\sqrt{\frac{s_u s_v}{t}} \|\Delta_{\widetilde{J}}\|_{\mathrm{F}} + \frac{5\delta^2}{\rho\lambda_r \sqrt{t}}, \quad (53)
\end{aligned}
$$

where we have used the generalized cone condition (52). Hence, we have the lower bound

$$\|\widehat{\Sigma}_x^{1/2} \Delta \widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} \geq \kappa_1 \|\Delta_{\widetilde{J}}\|_{\mathrm{F}} - \frac{\kappa_2 \delta^2}{\rho\lambda_r \sqrt{t}},$$

with

$$
\begin{aligned}
\kappa_1 &= \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u + t) \phi_{\min}^{\widehat{\Sigma}_y}(s_v + t)} - 9\sqrt{\frac{s_u s_v}{t}} \sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t) \phi_{\max}^{\widehat{\Sigma}_y}(t)}, \quad (54) \\
\kappa_2 &= 5\sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t) \phi_{\max}^{\widehat{\Sigma}_y}(t)}.
\end{aligned}
$$

Taking $t = c_1 s_u s_v$ for some sufficiently large constant $c_1 > 1$, with high probability, κ_1 can be lower bounded by a positive constant κ_0 only depending on M. To see this, note that by Lemma 6.5, (54) can be lower bounded by the difference of $\sqrt{M^{-1} - C\delta_u(2c_1 s_u s_v)}\sqrt{M^{-1} - C\delta_v(2c_1 s_u s_v)}$ and $9c_1^{-1/2}\sqrt{M + C\delta_u(c_1 s_u s_v)}\sqrt{M + C\delta_v(c_1 s_u s_v)}$, where δ_u and δ_v are defined as in Lemma 6.5. It is sufficient to show that $\delta_u(2c_1 s_u s_v)$, $\delta_v(2c_1 s_u s_v)$, $\delta_u(c_1 s_u s_v)$ and $\delta_v(c_1 s_u s_v)$ are sufficiently small to get a positive absolute constant κ_0. For the first term, when $2c_1 s_u s_v \leq p$, it is bounded by $\frac{2c_1 s_u s_v \log(ep)}{n}$ and is sufficiently small under the assumption (12). When $2c_1 s_u s_v > p$, it is bounded by $\frac{2c_1 s_u s_v}{n}$ and is also sufficiently small under (12). The same argument also holds for the other terms. Similarly, κ_2 can be upper bounded by some constant.

Together with (51), this brings the inequality

$$\|\Delta_{\widetilde{J}}\|_{\mathrm{F}}^2 \leq \frac{C_1 \sqrt{s_u s_v} \rho}{\lambda_r} \|\Delta_{\widetilde{J}}\|_{\mathrm{F}} + C_2 \left(\frac{\delta^2}{\lambda_r^2} + \left(\frac{\delta^2}{\rho\lambda_r \sqrt{t}} \right)^2 \right).$$

Solving this quadratic equation, we have

$$\|\Delta_{\widetilde{J}}\|_{\mathrm{F}}^2 \leq C \left(\frac{s_u s_v \rho^2}{\lambda_r^2} + \frac{\delta^2}{\lambda_r^2} + \left(\frac{\delta^2}{\rho\lambda_r \sqrt{t}} \right)^2 \right). \quad (55)$$

23

By (53), we have

$$\|\Delta_{\widetilde{j}^c}\|_{\mathrm{F}} \leq \sum_{k=2}^{K} \|\Delta_{J_k}\|_{\mathrm{F}} \leq 9\sqrt{\frac{s_u s_v}{t}}\|\Delta_{\widetilde{j}}\|_{\mathrm{F}} + \frac{5\delta^2}{\rho\lambda_r\sqrt{t}}. \tag{56}$$

Summing (55) and (56), we obtain a bound for $\|\Delta\|_{\mathrm{F}}$. According to Lemma 6.4, we may choose $\rho = \gamma\sqrt{\frac{\log(p+m)}{n}}$ for some large γ, so that (44) holds with high probability. By Lemma 6.1, $\delta \leq C\sqrt{\frac{r(s_u+s_v+\log(p+m))}{n}} \leq C'\rho\sqrt{t}$ with high probability. Hence,

$$\|\Delta\|_{\mathrm{F}} \leq C\sqrt{s_u s_v}\rho/\lambda_r, \tag{57}$$

with high probability. This completes the second step. Finally, by triangle inequality, we have $\|\widehat{A} - UV'\|_{\mathrm{F}} \leq \|\Delta\|_{\mathrm{F}} + \|\widetilde{A} - UV'\|_{\mathrm{F}}$. By (42) and (57), the proof is complete. $\qquad\square$

6.2 Proof of Theorem 4.2

Define

$$U^* = U\Lambda V'\Sigma_y\widehat{V}^{(0)}, \quad \Delta = \widehat{U}^{(1)} - U^*.$$

Lemma 6.7. *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*

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

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

Proof of Theorem 4.2. In the rest of this proof, we denote $\widehat{\Sigma}_x^{(1)}$, $\widehat{\Sigma}_y^{(1)}$ and $\widehat{\Sigma}_{xy}^{(1)}$ by $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$ and $\widehat{\Sigma}_{xy}$ for simplicity of notation. Note that they depends on $\mathcal{D}_1$, while the estimator $\widehat{V}^{(0)}$ depends on $\mathcal{D}_0$. Hence, $\widehat{V}^{(0)}$ is independent of the sample covariance matrices occurring in this proof. 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 last step, we derive the desired bound for $L(\widehat{U}, U)$.

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

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

After rearrangement, we have

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \leq \rho_u\sum_{j=1}^{p}\left[\|U_{j\cdot}^*\| - \|U_{j\cdot}^* + \Delta_{j\cdot}\|\right] + 2\,\mathsf{Tr}\left[\Delta'(\widehat{\Sigma}_{xy}\widehat{V}^{(0)} - \widehat{\Sigma}_x U^*)\right]. \tag{58}$$

24

For the first term on the right hand side of (58), we have

$$\sum_{j=1}^{p} \left(||U_{j\cdot}^*|| - ||U_{j\cdot}^* + \Delta_{j\cdot}|| \right) = \sum_{j \in S_u} ||U_{j\cdot}^*|| - \sum_{j \in S_u} ||U_{j\cdot}^* - \Delta_{j\cdot}|| - \sum_{j \in S_u^c} ||\Delta_{j\cdot}||$$

$$\leq \sum_{j \in S_u} ||\Delta_{j\cdot}|| - \sum_{j \in S_u^c} ||\Delta_{j\cdot}||.$$

For the second term on the right hand side of (58), we have

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

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

$$\rho_u \geq 4 \max_{1 \leq j \leq p} ||[\widehat{\Sigma}_{xy}\widehat{V}^{(0)} - \widehat{\Sigma}_x U^*]_{j\cdot}||, \tag{59}$$

we have

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \leq \frac{3\rho_u}{2} \sum_{j \in S_u} ||\Delta_{j\cdot}|| - \frac{\rho_u}{2} \sum_{j \in S_u^c} ||\Delta_{j\cdot}||. \tag{60}$$

Since $\sum_{j \in S_u} ||\Delta_{j\cdot}|| \leq \sqrt{s_u}\sqrt{\sum_{j \in S_u} ||\Delta_{j\cdot}||^2}$, (60) can be upper bounded by

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \leq \frac{3\sqrt{s_u}\rho_u}{2} \sqrt{\sum_{j \in S_u} ||\Delta_{j\cdot}||^2}. \tag{61}$$

This completes the first step.

Step 2. The inequality (60) implies the cone condition

$$\sum_{j \in S_u^c} ||\Delta_{j\cdot}|| \leq 3 \sum_{j \in S_u} ||\Delta_{j\cdot}||. \tag{62}$$

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 $\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) = \|n^{-1/2}X\Delta\|_{\mathrm{F}}^2$, where $X = [X_1, ..., X_n]' \in \mathbb{R}^{n \times p}$ denotes the data matrix. By triangle inequality, we have

$$\|n^{-1/2}X\Delta\|_{\mathrm{F}} \geq \|n^{-1/2}X\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}} - \sum_{k \geq 2} \|n^{-1/2}X\Delta_{J_{k*}}\|_{\mathrm{F}}$$

$$\geq \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u + t)}\|\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}} - \sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)} \sum_{k \geq 2} \|\Delta_{J_{k*}}\|_{\mathrm{F}},$$

25

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

$$
\begin{aligned}
\sum_{k\geq 2}\|\Delta_{J_k*}\|_{\mathrm{F}} &\leq \sqrt{t}\sum_{k\geq 2}\max_{j\in J_k}\|\Delta_{j\cdot}\| \leq \sqrt{t}\sum_{k\geq 2}\frac{1}{t}\sum_{j\in J_{k-1}}\|\Delta_{j\cdot}\| \qquad (63)\\
&\leq t^{-1/2}\sum_{j\in S_u^c}\|\Delta_{j\cdot}\| \leq 3t^{-1/2}\sum_{j\in S_u}\|\Delta_{j\cdot}\| \\
&\leq 3\sqrt{\frac{s_u}{t}}\sqrt{\sum_{j\in S_u}\|\Delta_{j\cdot}\|^2} \leq 3\sqrt{\frac{s_u}{t}}\|\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}}. \qquad (64)
\end{aligned}
$$

In the above derivation, we have used the construction of J_k and the cone condition (62). Hence, $\|n^{-1/2}X\Delta\|_{\mathrm{F}} \geq \kappa\|\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}}$ 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 $t = c_1 s_u$ 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 (61), we have

$$
\|\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}} \leq C\sqrt{s_u}\rho_u/(2\kappa_0^2). \qquad (65)
$$

By (63)-(64), we have

$$
\|\Delta_{(\widetilde{S}_u)^c*}\|_{\mathrm{F}} \leq \sum_{k\geq 2}\|\Delta_{J_k*}\|_{\mathrm{F}} \leq 3\sqrt{s_u/t}\|\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}} \leq 3c_1^{-1/2}\|\Delta_{\widetilde{S}_{u*}}\|_{\mathrm{F}}. \qquad (66)
$$

Summing (65) and (66), we have $\|\Delta\|_{\mathrm{F}} \leq C\sqrt{s_u}\rho$. By Lemma 6.7, we may choose $\rho_u \geq \gamma_u\sqrt{\frac{r+\log p}{n}}$ for some large γ_u so that (59) holds with high probability. Hence,

$$
\|\Delta\|_{\mathrm{F}} \leq C\sqrt{s_u(r+\log p)/n}, \qquad (67)
$$

with high probability. This completes the second step.

Step 3. Using the same argument in Step 2 of the proof of Theorem 3.2 (see supplementary material), we obtain the desired bound for $L(\widehat{U}, U)$. The proof is complete. $\qquad \square$

References

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

[2] S. Arora and B. Barak. *Computational complexity: a modern approach.* Cambridge University Press, 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] Z. Bao, J. Hu, G. Pan, and W. Zhou. Canonical correlation coefficients of high-dimensional normal vectors: finite rank case. *arXiv preprint arXiv:1407.7194*, 2014.

[5] 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.

[6] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[7] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[8] 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.

[9] A. Birnbaum, I.M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics*, 41(3):1055–1084, 2013.

[10] 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.

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

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

[13] P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

[14] J. Douglas and H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the AMS*, pages 421–439, 1956.

[15] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 655–664. ACM, 2013.

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

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

[18] B. Hajek, Y. Wu, and J. Xu. Computational lower bounds for community detection on random graphs. *arXiv preprint arXiv:1406.6625*, 2014.

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

[20] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[21] 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.

[22] I.M. Johnstone. Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy-Widom limits and rates of convergence. *Ann. Stat.*, 36:2638–2716, 2008.

[23] L. Le Cam. *Asymptotic Methods in Statistical Theory*. Springer-Verlag, 1986.

[24] 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.

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

[26] Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *arXiv preprint arXiv:1309.5914*, 2013.

[27] K. Mardia, J. Kent, and J. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[28] 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.

[29] B. Rossman. *Average-case complexity of detecting cliques*. PhD thesis, Massachusetts Institute of Technology, 2010.

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

[31] 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.

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

[33] T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *arXiv preprint arXiv:1408.5369*, 2014.

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

[35] Z. Wang, H. Lu, and H. Liu. Nonconvex statistical optimization: minimax-optimal sparse pca in polynomial time. *arXiv preprint arXiv:1408.5352*, 2014.

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

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

[38] 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.

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

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

Supplement to "Sparse CCA: Adaptive Estimation and Computational Barriers"

7 Computational Barriers for Sparse PCA

In this section, we show that the argument in Section 5 can be modified into a computational lower bound for sparse PCA under the Gaussian spiked covariance model. Being the most commonly used model for sparse PCA, the Gaussian spiked covariance model assumes that the data follows a multivariate Gaussian distribution $N_p(0, \Sigma)$, with

$$\Sigma = \Theta \Lambda \Theta' + I_p, \tag{68}$$

for some $\Theta \in O(p, r)$ and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_r)$ satisfying $\lambda_1 \geq ... \geq \lambda_r$. Let $\mathcal{Q}(n, s, p, r, \lambda; \kappa)$ denote the space of distributions of i.i.d. $\{X_i\}_{i=1}^n$ following $N_p(0, \Sigma)$, with Σ having the spiked covariance structure (68), where $|\text{supp}(\Theta)| \leq s$ and $\lambda \leq \lambda_r \leq \cdots \leq \lambda_1 \leq \kappa\lambda$. Here and after, we treat $\kappa \geq 1$ as an absolute constant that does not change with any other parameter. Recall that for any matrix A, P_A denotes the projection matrix onto its column subspace. The minimax estimation rate for Θ under the loss $\|P_{\widehat{\Theta}} - P_\Theta\|_F^2$ is

$$\frac{1}{n\lambda^2} s \left(r + \log \frac{ep}{s}\right). \tag{69}$$

See, for instance, [11]. However, to achieve the above minimax rate via computationally efficient methods such as those proposed in [9, 25, 7, 11, 35], researchers have required the sample size to satisfy

$$n \geq C \frac{s^2 \log p}{\lambda^2} \tag{70}$$

for some sufficiently large constant $C > 0$. When the condition (70) is violated, there is no known efficient algorithm even for consistent estimation for sparse PCA. Denote the $N_p(0, I_p)$ distribution by $\mathbb{Q}_0$. For i.i.d. observations $\{X_i\}_{i=1}^n$, A closely related sparse PCA testing problem is

$$H_0^P : \{X_i\}_{i=1}^n \sim \mathbb{Q}_0^n, \quad \text{v.s.} \quad H_1^P : \{X_i\}_{i=1}^n \sim \mathbb{Q} \in \mathcal{Q}(n, s, p, 1, \lambda; \kappa). \tag{71}$$

Berthet and Rigollet [6] showed that the assumption (70) is essentially necessary for all polynomial-time testing procedures if both the null and the alternative in (71) were enlarged to include all distributions that some tail probability bounds are satisfied. The same kind of enlargement of $\mathcal{Q}(n, s, p, 1, \lambda; \kappa)$ was also needed in the subsequent work [33] on estimation.

In the rest of this section, we show that the sample size condition (70) is essentially necessary for consistent sparse PCA estimation under the Gaussian spiked covariance model (68).

Hardness of sparse PCA To achieve complexity theoretic rigor, define the discretized sparse PCA probability space by

$$\mathcal{Q}^t(n, s, p, r, \lambda; \kappa) = \{\mathcal{L}([X]_t) : X \sim \mathbb{Q} \in \mathcal{Q}(n, s, p, r, \lambda; \kappa)\}.$$

The following theorem provides a computational lower bound for the sparse PCA estimation problem.

Theorem 7.1. *Suppose that Hypothesis A holds and that as $n \to \infty$, $2n \le p \le n^a$ for some constant $a > 1$, $n(\log n)^5 \le cs^4$ for some sufficiently small $c > 0$, and $\lambda = \frac{s^2}{2430n(\log(12n))^2}$. If for some $\delta \in (0, 2)$,*

$$\liminf_{n \to \infty} \frac{s^{2-\delta} \log p}{n\lambda^2} > 0,$$

then for any randomized polynomial-time estimator $\widehat{\theta}$,

$$\liminf_{n \to \infty} \sup_{\mathbb{Q} \in \mathcal{Q}^t(n,s,p,1,\lambda;3)} \mathbb{Q}\left\{\|P_{\widehat{\theta}} - P_\theta\|_{\mathrm{F}}^2 > \frac{1}{3}\right\} > \frac{1}{4}, \tag{72}$$

where the discretization level is $t = \lceil 4\log_2(p + n) \rceil$.

With the choice of n, s, p, λ and t in the theorem, Lemma 5.1 implies that the experiments $\mathcal{Q}(n, s, p, 1, \lambda; 3)$ and $\mathcal{Q}^t(n, s, p, 1, \lambda; 3)$ are asymptotically equivalent. Thus, the theorem states that for a sequence of asymptotically equivalent discretized sparse PCA model, the assumption (70) is necessary (up to a sub-polynomial factor) for any computationally efficient consistent estimator. On the other hand, under (70) applying a discretized version[1] of the efficient procedure in Section 3 of [11] on $\{[X_i]_t\}_{i=1}^n$ achieves the optimal rates (69).

A sketch of the reduction scheme In parallel to Section 5.2, we sketch below the reduction scheme omitting the discretization issue. A randomized polynomial-time reduction for the discretized model will be presented in Section 8.3, together with that for the sparse CCA problem.

The reduction for sparse PCA is a three-step procedure, where the first two steps are exactly the same as (31) – (33), the first two steps for sparse CCA reduction. Thus, after the first two steps, we have at hand $2n$ vectors $W_1, \ldots, W_{2n} \in \mathbb{R}^p$. Turn to the third step. For any estimation procedure, let $\widehat{\theta} = \widehat{\theta}(W_{n+1}, \ldots, W_{2n})$ be the resulting estimator by applying the procedure on the second half of the W_i's. We reject H_0^G in (22) if

$$\frac{\widehat{\theta}'\left(\frac{1}{n}\sum_{i=1}^n W_i W_i'\right)\widehat{\theta}}{\|\widehat{\theta}\|^2} \ge 1 + \frac{1}{4}k\eta_N. \tag{73}$$

---

[1]To be more precise, we need to replace the entries in the random matrix $\widetilde{Z}$ in Step 1 in Section 3.1 of [11] with discrete random variables sampled from the truncated dyadic approximation to the $N(0, 1)$ distribution spelled out below in Section 8.3.2. The constants used in the approximation can be chosen as $w = t$, $K = \lceil \log_2(3\sqrt{\log p}) \rceil$ and b as in (90) below.

The intuition behind the above scheme for sparse PCA is closely related to that of the reduction for sparse CCA. Recall that as discussed following $(31) - (35)$, when H_0^G holds, the W_i's are close in total variation to $2n$ i.i.d. random vectors from the $N_p(0, I_p)$ distribution, while when H_1^G holds, the joint distribution is close to a mixture of the distribution of $2n$ i.i.d. random vectors following the Gaussian spiked covariance model in (36). Thus, following the same intuition as discussed after (35), the behavior of the LHS of (73) is similar to that of the LHS of (35), either under H_0^G or H_1^G, which leads to the following counterpart of Theorem 5.2.

Theorem 7.2. *For some sufficiently small constant $c > 0$, assume $\frac{k^2}{N(\log N)^2} \vee \frac{N(\log N)^5}{k^4} \leq c$, $cN \leq n \leq N/12$ and $p \geq 2n$. Then, for any $\widehat{\theta}$ such that*

$$\sup_{\mathbb{Q} \in \mathcal{Q}(n, 3k/2, p, 1, k\eta_N/2; 3)} \mathbb{Q} \left\{ \|P_{\widehat{\theta}} - P_\theta\|_{\mathrm{F}}^2 > \frac{1}{3} \right\} \leq \beta,$$

the test ψ defined by $(31) - (33)$ and (73) satisfies

$$\mathbb{P}_{H_0^G} \psi + \mathbb{P}_{H_1^G} (1 - \psi) < \beta + \frac{4n}{N} + C(n^{-1} + N^{-1} + e^{-C'k}),$$

for sufficiently large n with some constants $C, C' > 0$.

Remark 7.1. With slight modification, the above reduction scheme leads to a computational lower bound for the sparse PCA testing problem (71). If we have a testing procedure for (71), we can simply replace the third step (73) with directly applying the test to $\{W_i\}_{i=1}^n$ and then using the output of this test as the testing result for (22). A simple modification of the proof of Theorem 7.1 then leads to a comparable computational lower bound for sparse PCA testing.

8 Proofs for Computational Lower Bounds

In this section, we prove the results stated in Sections 5 and 7. The proof of Lemma 5.2 is given in Section 8.1. In Section 8.2, we prove Theorems 5.2 and 7.2. These results and proofs do not consider the issue of discretization. The main purpose is to help the readers get the intuition behind the problem without worrying about rigor at the theoretical computer science level. A rigorous treatment of the computational lower bounds is presented in Section 8.3, where we first prove Lemma 5.1 on asymptotically equivalent discretized models, and then show how the reductions for the continuous Gaussian models in Sections 5 and 7 can be made into truly randomized polynomial-time reductions. Discretized versions of Theorems 5.2 and 7.2 are presented, followed by the proofs of Theorems 5.1 and 7.1.

8.1 Proof of Lemma 5.2

We first verify that $(29) - (30)$ are proper density functions when $|\mu| \leq 3\sqrt{\eta_N \log N}$, which is a direct consequence of the following lemma.

Lemma 8.1. *For any $k \leq N/12$, $|\mu| \leq 3\sqrt{\eta_N \log N}$ and $|x| \leq 3\sqrt{\log N}$, we have*

$$\delta_N^{-1} \left| \bar{\phi}_\mu(x) - \phi_0(x) \right| \leq \frac{4}{5} \phi_0(x).$$

Proof. By definition,

$$\delta_N^{-1} \left| \bar{\phi}_\mu(x) - \phi_0(x) \right| = (2\delta_N)^{-1} \phi_0(x) \left| \exp\left(\mu x - \frac{\mu^2}{2} \right) + \exp\left(-\mu x - \frac{\mu^2}{2} \right) - 2 \right|.$$

Under the conditions of the lemma, we have $|\mu x| + \frac{\mu^2}{2} \leq \frac{1}{2}$, which implies that

$$\left| \exp\left(\mu x - \frac{\mu^2}{2} \right) + \exp\left(-\mu x - \frac{\mu^2}{2} \right) - 2 \right| \leq \mu^2 + \frac{4}{3} |\mu x|^2 + \frac{\mu^4}{3} \leq 8\mu^2 \log N.$$

We complete the proof by combining the last two displays. $\qquad\square$

By the result of the above lemma and the definitions of $f_{\mu,0}$ and $f_{\mu,1}$, we immediately have $f_{\mu,0} \geq 0$ and $f_{\mu,1} \geq 0$. Hence, they are valid density functions. The following lemma further controls the rescaling constants in (29) and (30).

Lemma 8.2. *There exists an absolute constant $C > 0$ such that for any $|\mu| \leq 1$, $|M_i - 1| \leq CN^{-4}$ for $i = 0, 1$.*

Proof. Note that

$$
\begin{aligned}
1 = \int f_{\mu,0}(x) \mathrm{d}x &= M_0 \int \left(\phi_0(x) - \delta_N^{-1}(\bar{\phi}_\mu(x) - \phi_0(x)) \right) \mathrm{d}x \\
&\quad - M_0 \int_{|x| > 3\sqrt{\log N}} \left(\phi_0(x) - \delta_N^{-1}(\bar{\phi}_\mu(x) - \phi_0(x)) \right) \mathrm{d}x \\
&= M_0 - M_0 \int_{|x| > 3\sqrt{\log N}} \left(\phi_0(x) - \delta_N^{-1}(\bar{\phi}_\mu(x) - \phi_0(x)) \right) \mathrm{d}x.
\end{aligned}
$$

The integral on the RHS is upper bounded by

$$\left(1 + \delta_N^{-1} \right) \int_{|x| > 3\sqrt{\log N}} \phi_0(x) \mathrm{d}x + \delta_N^{-1} \int_{|x| > 3\sqrt{\log N}} \bar{\phi}_\mu(x) \mathrm{d}x \leq CN^{-4},$$

where the last inequality comes from standard Gaussian tail bounds. This readily implies $|M_0 - 1| \leq CN^{-4}$ The desired bound on M_1 follows from similar arguments. $\qquad\square$

Proof of Lemma 5.2. Define

$$g_i(x) = \phi_0(x) - (-1)^{i+1} \delta_N^{-1}(\bar{\phi}_\mu(x) - \phi_0(x)), \quad \text{for } i = 0, 1.$$

Then we have for $i = 0$ and 1,

$$f_{\mu,i}(x) = g_i(x) - (1 - M_i \mathbf{1}_{\{|x| \leq 3\sqrt{\log N}\}}) g_i(x),$$

By Lemma 8.1 and Lemma 8.2,

$$\int |f_{\mu,i}(x) - g_i(x)|\mathrm{d}x \le \int \left|(1 - M_i \mathbf{1}_{\{|x|\le 3\sqrt{\log N}\}})g_i(x)\right|\mathrm{d}x$$

$$\le |1 - M_i| \int |g_i(x)|dx + M_i \int_{|x|>3\sqrt{\log N}} |g_i(x)|\mathrm{d}x \le CN^{-3}. \qquad (74)$$

Therefore, we have

$$\mathsf{TV}\left(\frac{1}{2}(f_{\mu,0} + f_{\mu,1}), \phi_0\right) = \frac{1}{2}\int \left|\phi_0(x) - \frac{1}{2}(f_{\mu,0}(x) + f_{\mu,1}(x))\right|\mathrm{d}x$$

$$\le \frac{1}{2}\int \left|\phi_0(x) - \frac{1}{2}(g_0(x) + g_1(x))\right|\mathrm{d}x + \frac{1}{4}\sum_{i=0,1}\int |f_{\mu,i}(x) - g_i(x)|\mathrm{d}x$$

$$\le CN^{-3},$$

where the last inequality is due to the identity $\phi_0 = \frac{1}{2}(g_0 + g_1)$ and (74). In addition, we have

$$\mathsf{TV}\left(\delta_N f_{\mu,1} + (1 - \delta_N)\frac{1}{2}(f_{\mu,0} + f_{\mu,1}), \bar{\phi}_\mu\right)$$

$$= \frac{1}{2}\int \left|\delta_N f_{\mu,1}(x) + \frac{1 - \delta_N}{2}(f_{\mu,0}(x) + f_{\mu,1}(x)) - \bar{\phi}_\mu(x)\right|\mathrm{d}x$$

$$\le \frac{1}{2}\int \left|\delta_N g_1(x) + \frac{1 - \delta_N}{2}(g_0(x) + g_1(x)) - \bar{\phi}_\mu(x)\right|\mathrm{d}x$$

$$+ \sum_{i=0,1}\frac{1 - (-1)^i\delta_N}{4}\int |f_{\mu,i}(x) - g_i(x)|dx$$

$$\le CN^{-3}.$$

Here, the last inequality is due to the identity $\delta_N g_1 + \frac{1-\delta_N}{2}(g_0 + g_1) = \bar{\phi}_\mu$ and (74). This completes the proof. $\qquad\square$

8.2   Proofs of Theorems 5.2 and 7.2

To facilitate the proof, we first state and prove two lemmas which characterize the distributions of the W_i's and the (X_i, Y_i)'s under H_0^G and H_1^G respectively.

Let $\mathcal{L}(\{W_i\}_{i=1}^{2n})$ denote the joint distribution of $\{W_i\}_{i=1}^{2n}$ and $\mathcal{L}(\{(X_i, Y_i)\}_{i=1}^n, \{W_i\}_{i=1}^n)$ that of $\{(X_i, Y_i)\}_{i=1}^n$ and $\{W_i\}_{i=1}^n$. In addition, denote the $(p+m)$-dimensional normal distribution with mean zero and covariance (37) by $\mathbb{P}_{\theta,\tau}$, and the p-dimensional normal distribution with mean zero and covariance (36) by $\mathbb{Q}_{\theta,\tau}$. When $\tau = 0$, the two distributions reduce to $N_{p+m}(0, I_{p+m})$ and $N_p(0, I_p)$, which are denoted by $\mathbb{P}_0$ and $\mathbb{Q}_0$, respectively. We use $\mathbb{P} \times \mathbb{Q}$ to denote the product measure of two probability measures $\mathbb{P}$ and $\mathbb{Q}$.

The first lemma concerns the joint distributions of the W_i, X_i and Y_i vectors when H_0^G holds. Roughly speaking, under H_0^G, the joint distribution of $\{W_i\}_{i=1}^{2n}$ is close in total variation to that of a random sample of size $2n$ from $\mathbb{Q}_0$, and that of $\{(X_i, Y_i)\}_{i=1}^n$ and $\{W_i\}_{i=1}^n$

is close in total variation to that of a random sample of size n from $\mathbb{P}_0$ together with an independent random sample of size n from $\mathbb{Q}_0$.

Lemma 8.3. *Suppose $A \sim \mathcal{G}(N, 1/2)$. There exists an absolute constant $C > 0$ such that*

$$\mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{2n}), \mathbb{Q}_0^{2n}) \leq CN^{-1},$$
$$\mathsf{TV}\left(\mathcal{L}\left(\{(X_i, Y_i)\}_{i=1}^{n}, \{W_i\}_{i=1}^{n}\right), \mathbb{P}_0^n \times \mathbb{Q}_0^n\right) \leq CN^{-1}.$$

Proof. Recall η_N defined in (27) and $h_{\mu,0}$ in Lemma 5.2. Let ν be $N(0, \eta_N)$, and $\bar{\nu}$ be the distribution obtained by restricting ν on the set $[-3\sqrt{\eta_N \log N}, 3\sqrt{\eta_N \log N}]$. Then the μ_i's in (31) are i.i.d. r.v.'s following the distribution $\bar{\nu}$.

For each $i \in [2n]$ and each $j \in [2n]$, define i.i.d. random variables $\overline{W}_{ij} \sim N(0, 1)$. For each $i \in [2n]$ and $2n < j \leq p$, define $\overline{W}_{ij} = W_{ij}$. Let $\overline{W}_i = (\overline{W}_{i1}, ..., \overline{W}_{ip})'$. We also define

$$\overline{X}_i = \frac{1}{\sqrt{2}}(\overline{W}_{n+i} + Z_i), \quad \overline{Y}_i = \frac{1}{\sqrt{2}}(\overline{W}_{n+i} - Z_i), \tag{75}$$

for all $i \in [2n]$, where the Z_i's are the same random vectors as in (34). It is straightforward to verify that $\mathcal{L}(\{\overline{W}_i\}_{i=1}^{2n}) = \mathbb{Q}_0^{2n}$ and $\mathcal{L}\left(\{(\overline{X}_i, \overline{Y}_i)\}_{i=1}^{n}, \{\overline{W}_i\}_{i=1}^{n}\right) = \mathbb{P}_0^n \times \mathbb{Q}_0^n$. By the data-processing inequality, we have

$$\mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{2n}), \mathbb{Q}_0^{2n}) \leq \mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{n}), \mathcal{L}(\{\overline{W}_i\}_{i=1}^{n})),$$
$$\mathsf{TV}\left(\mathcal{L}\left(\{(X_i, Y_i)\}_{i=1}^{n}, \{W_i\}_{i=1}^{n}\right), \mathbb{P}_0^n \times \mathbb{Q}_0^n\right) \leq \mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{n}), \mathcal{L}(\{\overline{W}_i\}_{i=1}^{n})).$$

Hence, it is sufficient to bound $\mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{n}), \mathcal{L}(\{\overline{W}_i\}_{i=1}^{n}))$. Conditioning on μ_i, W_{ij} follows $h_{\mu_i,0}$ when $A \sim \mathcal{G}(N, k)$. Therefore,

$$\mathsf{TV}(W_{ij}, \overline{W}_{ij}) = \mathsf{TV}(\int h_{\mu_i,0} \mathrm{d}\bar{\nu}(\mu_i), \phi_0)$$
$$\leq \sup_{|\mu_i| \leq 3\sqrt{\eta_N \log N}} \mathsf{TV}(h_{\mu_i,0}, \phi_0) \leq CN^{-3}.$$

Here the last inequality is due to Lemma 5.2. Applying Lemma 7 of [26], we obtain

$$\mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{n}), \mathcal{L}(\{\overline{W}_i\}_{i=1}^{n})) \leq \sum_{i=1}^{n} \sum_{j=1}^{n} \mathsf{TV}(W_{ij}, \overline{W}_{ij}) \leq CN^{-1}.$$

This completes the proof. $\qquad\square$

The second lemma characterizes the joint distributions of the W_i, X_i and Y_i vectors when H_1^G holds. In this case, the joint distribution $\{W_i\}_{i=1}^{2n}$ is close in total variation to a mixture of the joint distribution of a random sample of size $2n$ from $\mathbb{Q}_{\theta,\tau}$, and that of $\{(X_i, Y_i)\}_{i=1}^{n}$ and $\{W_i\}_{i=1}^{n}$ is close in total variation to that of a mixture over the joint distribution of a random sample of size n from $\mathbb{P}_{\theta,\tau}$ together with an independent random sample of size n from $\mathbb{Q}_{\theta,\tau}$ with the same θ and τ parameters. Here, the mixture is defined by a prior distribution π on the (θ, τ) pair, which is supported on a region where θ is sparse and τ is bounded

35

away from zero. For notational convenience, for any distribution $\mathbb{P}_\beta$ indexed by parameter $\beta \in B$ and any probability measure π on B, we let $\int \mathbb{P}_\beta d\nu(\beta)$ denote the probability measure $\mathbb{P}$ defined by $\mathbb{P}(E) = \int \mathbb{P}_\beta(E) d\nu(\beta)$ for any event E. When $\beta \sim \nu$ is a random variable and $\mathbb{P}_\beta = \mathcal{L}(W|\beta)$ is the conditional distribution of $W|\beta$, we also write $\int \mathcal{L}(W|\beta) d\nu(\beta)$ to represent the marginal distribution of W after integrating out β.

Lemma 8.4. *Suppose $A \sim \mathcal{G}(N, 1/2, k)$. There exists a distribution π supported on the set*

$$\left\{ (\theta, \tau) : \theta \in S^{p-1}, |\mathrm{supp}(\theta)| \le 3k/2, \tau \in [k\eta_N/2, 3k\eta_N/2] \right\}, \tag{76}$$

such that for some absolute constants $C_1, C_2 > 0$,

$$\mathsf{TV}(\mathcal{L}(\{W_i\}_{i=1}^{2n}), \int \mathbb{Q}_{\theta,\tau}^{2n} d\pi(\theta, \tau)) \le C_1 \left(e^{-C_2 k} + \frac{1}{N} \right) + \frac{4n}{N},$$

$$\mathsf{TV}(\mathcal{L}(\{(X_i, Y_i)\}_{i=1}^n, \{W_i\}_{i=1}^n), \int (\mathbb{P}_{\theta,\tau}^n \times \mathbb{Q}_{\theta,\tau}^n) d\pi(\theta, \tau)) \le C_1 \left(e^{-C_2 k} + \frac{1}{N} \right) + \frac{4n}{N}.$$

Proof. Recall η_N defined in (27) and $h_{\mu,0}$ and $h_{\mu,1}$ defined in Lemma 5.2. As in the proof of Lemma 8.3, let ν be $N(0, \eta_N)$, and $\bar{\nu}$ the distribution obtained by restricting ν on the set $[-3\sqrt{\eta_N \log N}, 3\sqrt{\eta_N \log N}]$. Then the μ_i's in (31) are i.i.d. r.v.'s following $\bar{\nu}$. Simple calculus shows that $\int \phi_0(x) d\nu(\mu) = \phi_0(x)$ is the density function of $N(0,1)$, and $\int \bar{\phi}_\mu(x) d\nu(\mu)$ gives the density function of $N(0, 1 + \eta_N)$.

We first focus on the case $p = 2n$. The case of $p \ge 2n$ will be treated at the end of the proof.

Recall that $(\epsilon_1, ..., \epsilon_{2n})$ are the indicators of the rows of A_0 whether the corresponding vertices belong to the planted clique, and $(\gamma_1, ..., \gamma_p)$ are the corresponding indicators of the columns of A_0. Let $(\widetilde{\epsilon}_1, ..., \widetilde{\epsilon}_{2n})$ and $(\widetilde{\gamma}_1, ..., \widetilde{\gamma}_p)$ be i.i.d. Bernoulli random variables with mean $\delta_N = k/N$. Define a matrix $\widetilde{A}_0$, where an entry $(\widetilde{A}_0)_{ij} = 1$ if $\widetilde{\epsilon}_i = \widetilde{\gamma}_j = 1$ and is an independent instantiation of the Bernoulli(1/2) distribution otherwise. Then, we define $\widetilde{W}$ with entries

$$\widetilde{W}_{ij} = (B_0)_{ij}(1 - (\widetilde{A}_0)_{ij}) + (B_1)_{ij}(\widetilde{A}_0)_{ij}.$$

Then, by Theorem 4 of [13] and the data-processing inequality, we have

$$\mathsf{TV}(\mathcal{L}(\widetilde{W}), \mathcal{L}(W)) \le \mathsf{TV}(\mathcal{L}(\widetilde{\epsilon}, \widetilde{\gamma}), \mathcal{L}(\epsilon, \gamma)) \le \frac{4n}{N}. \tag{77}$$

Recall $h_{\mu,0}$ and $h_{\mu,1}$ defined in Lemma 5.2. By the definition of $\widetilde{W}$, conditioning on μ_i and $\widetilde{\gamma}_j = 0$, $\widetilde{W}_{ij} \sim h_{\mu_i,0}$, while conditioning on μ_i and $\widetilde{\gamma}_j = 1$, $\widetilde{W}_{ij} \sim h_{\mu_i,1}$.

Further define $\overline{W}_{ij}$ by setting

$$\overline{W}_{ij}|(\widetilde{\gamma}_j = 0, \mu_i) \sim \phi_0, \quad \overline{W}_{ij}|(\widetilde{\gamma}_j = 1, \mu_i) \sim \bar{\phi}_{\mu_i},$$

where $\bar{\phi}_{\mu_i}$ is defined according to (28). By Lemma 5.2 and Lemma 7 of [26], uniformly over $\max_i |\mu_i| \le 3\sqrt{\eta_N \log N}$, we have

$$\mathsf{TV}\left(\mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu), \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu) \right) \le \sum_{i=1}^{2n} \sum_{j=1}^p \mathsf{TV}\left(\mathcal{L}(\widetilde{W}_{ij}|\widetilde{\gamma}_j, \mu_i), \mathcal{L}(\overline{W}_{ij}|\widetilde{\gamma}_j, \mu_i) \right) \le C N^{-1}$$

36

for some constant $C > 0$.

Next, we integrate the above bound over μ. To this end, first note that

$$\mathsf{TV}(\nu, \bar{\nu}) = \int_{|\mu| > 3\sqrt{\eta_N \log N}} \mathrm{d}\nu(\mu) = \int_{|x| > 3\sqrt{\log N}} \phi_0(x)\mathrm{d}x \leq CN^{-4}.$$

With slight abuse of notation, let $\int \mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu)$ (resp. $\int \mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu)\mathrm{d}\nu(\mu)$) denote the conditional distribution of $\widetilde{W}|\widetilde{\gamma}$ if the coordinates of $\mu = (\mu_1, \ldots, \mu_{2n})$ were i.i.d. following $\bar{\nu}$ (resp. ν), and let $\int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu)$ and $\int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\nu(\mu)$ be analogously defined. Then, conditioning on $\widetilde{\gamma}$, we obtain

$$\mathsf{TV}(\int \mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu), \int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\nu(\mu))$$

$$\leq \quad \mathsf{TV}(\int \mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu), \int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu))$$

$$+ \mathsf{TV}(\int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu), \int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\nu(\mu))$$

$$\leq \quad \sup_{\max_i |\mu_i| \leq 3\sqrt{\eta_N \log N}} \mathsf{TV}(\mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu), \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)) + Cn\mathsf{TV}(\bar{\nu}, \nu)$$

$$\leq \quad CN^{-1}.$$

Here, the first inequality comes from the triangle inequality, the second from the definition of total variation distance. For each given $\widetilde{\gamma} = (\widetilde{\gamma}_1, ..., \widetilde{\gamma}_n)$, define $s = \sum_{j=1}^n \widetilde{\gamma}_j = \sum_{j=1}^n \widetilde{\gamma}_j^2 = \|\widetilde{\gamma}_j\|^2$, $\theta = s^{-1/2}\widetilde{\gamma}$ and $\tau = s\eta_N$. Note that both θ and τ are functions of $\widetilde{\gamma}$. Then observe that

$$\int \mathcal{L}(\overline{W}|\widetilde{\gamma}, \mu)\mathrm{d}\nu(\mu) = \mathbb{Q}_{\theta,\tau}^{2n},$$

which implies for $\mathcal{L}(\widetilde{W}|\widetilde{\gamma}) = \int \mathcal{L}(\widetilde{W}|\widetilde{\gamma}, \mu)\mathrm{d}\bar{\nu}(\mu)$,

$$\mathsf{TV}\left(\mathcal{L}(\widetilde{W}|\widetilde{\gamma}), \mathbb{Q}_{\theta,\tau}^{2n}\right) \leq CN^{-1}.$$

Define the event

$$Q = \{\widetilde{\gamma} : |s - k| \leq k/2\}.$$

Then, by Bernstein's inequality, $\mathbb{P}(Q^c) \leq e^{-Ck}$. Let $\widetilde{\pi}$ be the joint distribution of (θ, τ), and π be the distribution obtained from renormalizing the restriction of $\widetilde{\pi}$ on $\{(\theta(\widetilde{\gamma}), \tau(\widetilde{\gamma})) : \widetilde{\gamma} \in Q\}$ which is exactly the set in (76). Then we have $\mathsf{TV}(\pi, \widetilde{\pi}) \leq C\mathbb{P}(Q^c) \leq Ce^{-Ck}$. In addition, we note that $\mathcal{L}(\widetilde{W}|\widetilde{\gamma}) = \mathcal{L}(\widetilde{W}|\theta, \tau)$ since there exists one-to-one identification between the pair (θ, τ) and $\widetilde{\gamma}$. Therefore, we have

$$\mathsf{TV}(\mathcal{L}(\widetilde{W}), \int \mathbb{Q}_{\theta,\tau}^{2n}\mathrm{d}\pi(\theta, \tau)) \quad \leq \quad \mathsf{TV}(\mathcal{L}(\widetilde{W}), \int \mathcal{L}(\widetilde{W}|\theta, \tau)\mathrm{d}\pi(\theta, \tau))$$

$$+ \mathsf{TV}(\int \mathcal{L}(\widetilde{W}|\theta, \tau)\mathrm{d}\pi(\theta, \tau), \int \mathbb{Q}_{\theta,\tau}^{2n}\mathrm{d}\pi(\theta, \tau))$$

$$\leq \quad \mathsf{TV}(\widetilde{\pi}, \pi) + \sup_{\theta,\tau} \mathsf{TV}(\mathcal{L}(\widetilde{W}|\theta, \tau), \mathbb{Q}_{\theta,\tau}^{2n})$$

$$\leq \quad C\left(e^{-Ck} + N^{-1}\right).$$

37

Here, the second inequality holds since $\mathcal{L}(\widetilde{W}) = \int \mathcal{L}(\widetilde{W}|\theta,\tau)\mathrm{d}\widetilde{\pi}(\theta,\tau)$. Hence, by (77),

$$\mathsf{TV}(\mathcal{L}(W), \int \mathbb{Q}_{\theta,\tau}^{2n}\mathrm{d}\pi(\theta,\tau)) \le C\left(e^{-Ck} + N^{-1}\right) + \frac{4n}{N}.$$

Note that on the support of π, the parameter (θ,τ) belongs to the set (76). An application of data-processing inequality leads to the conclusion. When $p \ge 2n$, we may first analyze the distribution of the first $2n$ coordinates using the above arguments. The remaining $2n - p$ coordinates are exact, and the total variation bound is zero. This establishes the first inequality. The second inequality is a direct consequence of the data processing inequality. This completes the proof. $\qquad\square$

Proof of Theorem 5.2. By (36) and (37), we have $u = \theta/\sqrt{\tau/2+1}$, and so $\widehat{\theta} = \widehat{u}/\|\widehat{u}\|$ can be viewed as an estimator for θ. Abbreviate $\frac{1}{n}\sum_{i=1}^n W_i W_i'$ by $\widehat{\Sigma}$. We can rewrite the testing function ψ as

$$\psi(X, Y, W) = \psi(A, \mu, B_0, B_1, Z) = \mathbf{1}\left\{\widehat{\theta}'\widehat{\Sigma}\widehat{\theta} \ge 1 + k\eta_N/4\right\}.$$

Here, $\mu = (\mu_1, \ldots, \mu_{2n})$ collects the random variables in (31) and $Z = [Z_1', \ldots, Z_n']'$ consists of the random vectors used in defining the (X_i, Y_i)'s in (34). Thus, it is clear that ψ is a randomized test for the Planted Clique detection problem (22).

Note that for any (θ, τ) in the support of π, we have

$$\begin{aligned}
\mathbb{Q}_{\theta,\tau}^n &\in \mathcal{Q}\left(n, \frac{3k}{2}, p, 1, \frac{k\eta_N}{2}; 3\right) \quad \text{and} \\
\mathbb{P}_{\theta,\tau}^n &\in \mathcal{P}\left(n, \frac{3k}{2}, \frac{3k}{2}, p, p, 1, \frac{k\eta_N}{8}; 4\right),
\end{aligned} \tag{78}$$

where the second relation holds when $\frac{k^2}{180N(\log N)^2} \le 1$ which is satisfied under the condition of the theorem. To simplify notation, we denote below $\mathbb{P}_0^n \times \mathbb{Q}_0^n$ and $\mathbb{P}_{\theta,\tau}^n \times \mathbb{Q}_{\theta,\tau}^n$ by $\mathbb{P}_0^{\mathrm{joint}}$ and $\mathbb{P}_{\theta,\tau}^{\mathrm{joint}}$, respectively.

We now bound the testing errors of (38). For Type-I error, Lemma 8.3 implies

$$\mathbb{P}_{H_0^G}\psi \le \mathbb{P}_0^{\mathrm{joint}}\psi + CN^{-1}.$$

Note that under $\mathbb{P}_0^{\mathrm{joint}}$, $\widehat{\theta}$ and $\widehat{\Sigma}$ are independent. Conditioning on $\widehat{\theta}$ and using Bernstein's inequality, we have

$$\widehat{\theta}'\widehat{\Sigma}\widehat{\theta} = 1 + \frac{1}{n}\sum_{i=1}^n \left(|\widehat{\theta}'W_i|^2 - \|\widehat{\theta}\|^2\right) > 1 + \frac{k\eta_N}{4},$$

with probability at most $\exp\left(-\frac{Cnk^4}{N^2(\log N)^4}\right)$. Integrating over $\widehat{\theta}$, we have

$$\mathbb{P}_{H_0^G}\psi \le \exp\left(-\frac{Cnk^4}{N^2(\log N)^4}\right) + CN^{-1} \le C(n^{-1} + N^{-1}), \tag{79}$$

where the last inequality holds under the assumptions $\frac{N(\log N)^5}{k^4} \leq c$ and $cN \leq n \leq N/12$ for some sufficiently small constant $c > 0$.

Turn to the Type-II error. Lemma 8.4 implies

$$\mathbb{P}_{H_1^G}(1 - \psi) \leq \mathbb{P}_\pi^{\text{joint}}(1 - \psi) + C\left(e^{-Ck} + N^{-1}\right) + \frac{4n}{N}, \tag{80}$$

where we have used the notation $\mathbb{P}_\pi^{\text{joint}} = \int \mathbb{P}_{\theta,\tau}^{\text{joint}} d\pi$. For each $\mathbb{P}_{\theta,\tau}^{\text{joint}}$ in the support of π, W_i has representation

$$W_i = \sqrt{\tau} g_i \theta + \epsilon_i,$$

where the g_i's and the ϵ_i's are independently distributed according to $N(0,1)$ and $N_p(0, I_p)$, and are independent across $i = 1, ..., 2n$, and $\tau \geq k\eta_N/2$. Therefore,

$$\widehat{\theta}'\widehat{\Sigma}\widehat{\theta} = \tau|\widehat{\theta}'\theta|^2 \left(\frac{1}{n}\sum_{i=1}^n g_i^2\right) + \frac{1}{n}\sum_{i=1}^n |\widehat{\theta}'\epsilon_i|^2 + \frac{2\sqrt{\tau}}{n}\widehat{\theta}'\theta\sum_{i=1}^n g_i\epsilon_i'\widehat{\theta}.$$

After rearrangement, we have

$$\left|\widehat{\theta}'\widehat{\Sigma}\widehat{\theta} - (1 + \tau)\right| \leq \left|\frac{1}{n}\sum_{i=1}^n (g_i^2 - 1)\right| + \tau\min\{|(\widehat{\theta} - \theta)'\theta|^2, |(\widehat{\theta} + \theta)'\theta|^2\}$$
$$+ \left|\frac{1}{n}\sum_{i=1}^n (|\widehat{\theta}'\epsilon_i|^2 - 1)\right| + \left|\frac{2}{n}\sum_{i=1}^n g_i(\epsilon_i'\widehat{\theta})\right|,$$

where $\min\{|(\widehat{\theta} - \theta)'\theta|^2, |(\widehat{\theta} + \theta)'\theta|^2\}$ is bounded by

$$\min\left\{||\widehat{\theta} - \theta||^2, ||\widehat{\theta} + \theta||^2\right\} \leq \frac{4\min\{||\widehat{u} - u||^2, ||\widehat{u} + u||^2\}}{||u||^2}$$
$$\leq \frac{4}{\sigma_{\min}^2(\Sigma_x)||u||^2}L(\widehat{u}, u) \leq 4\frac{\sigma_{\max}^2(\Sigma_x)}{\sigma_{\min}^2(\Sigma_x)}L(\widehat{u}, u)$$
$$\leq 32^2 L(\widehat{u}, u). \tag{81}$$

Here, Σ_x is defined in (37) and the last inequality is due to (78). Together with (38), the above bound implies that for each (θ, τ) pair in the support of π,

$$\mathbb{P}_{\theta,\tau}^{\text{joint}}\left\{\min\{|(\widehat{\theta} - \theta)'\theta|^2, |(\widehat{\theta} + \theta)'\theta|^2\} > \frac{1}{3}\right\} \leq \beta. \tag{82}$$

By Bernstein's inequality, we have

$$\mathbb{P}_{\theta,\tau}^{\text{joint}}\left\{\left|\frac{1}{n}\sum_{i=1}^n (g_i^2 - 1)\right| + \left|\frac{1}{n}\sum_{i=1}^n (|\widehat{\theta}'\epsilon_i|^2 - 1)\right| + \left|\frac{2}{n}\sum_{i=1}^n g_i(\epsilon_i'\widehat{\theta})\right| > C\sqrt{\frac{\log n}{n}}\right\} \leq n^{-C'}.$$

Combining the above analysis and using the assumptions that $\frac{N(\log N)^5}{k^4} \leq c$ and $cN \leq n \leq N/12$, we have

$$\mathbb{P}_{\theta,\tau}^{\text{joint}}(1 - \psi) \leq \beta + n^{-C'}. \tag{83}$$

39

Integrating over (θ, τ) according to the prior π and applying (80), we obtain

$$\mathbb{P}_{H_1^G}(1 - \psi) \leq \beta + n^{-C'} + C\left(e^{-Ck} + N^{-1}\right) + \frac{4n}{N}.$$

Summing up the Type-I and Type-II errors, we have

$$\mathbb{P}_{H_0^G}\psi + \mathbb{P}_{H_1^G}(1 - \psi) \leq \beta + \frac{4n}{N} + C(n^{-1} + N^{-1} + e^{-C'k}). \tag{84}$$

Thus, the proof is complete. $\qquad\square$

Proof of Theorem 7.2. The proof is similar to that of Theorem 5.2. Without loss of generality, assume $\|\widehat{\theta}\| = 1$. Then, the proof of Theorem 5.2 essentially follows while the major difference is that the inequality (81) is replaced by

$$\begin{aligned}
\min\{|(\widehat{\theta} - \theta)'\theta|^2, |(\widehat{\theta} + \theta)'\theta|^2\} &\leq \min\left\{\|\widehat{\theta} - \theta\|^2, \|\widehat{\theta} + \theta\|^2\right\} \\
&\leq \|P_{\widehat{\theta}} - P_\theta\|_F^2.
\end{aligned}$$

With the above modification, the key inequalities (79), (80), (82), (83) and hence the conclusion (84) all follow. This completes the proof. $\qquad\square$

8.3 Discretized models and proofs of Theorems 5.1 and 7.1

In this section, we prove the rigorous computational lower bounds in Theorems 5.1 and 7.1. To state these results, we have adopted the asymptotically equivalent discretization framework established in [26] in Section 5. In what follows, we first prove Lemma 5.1 which bounds the Le Cam distance between multivariate Gaussian experiments and their discretized versions. Then we describe how the reduction for continuous models introduced in Section 5.2 and Section 7 can be slightly modified to become truly randomized polynomial-time reductions connecting the Planted Clique problem (22) and the discrete sparse CCA and sparse PCA estimation problems. Proofs of Theorems 5.1 and 7.1 then follow.

8.3.1 Proof of Lemma 5.1

Recall that for two statistical experiments $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ and $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$, the Le Cam deficiency of $\mathcal{P}$ with respect to $\mathcal{Q}$ is defined by $\delta(\mathcal{P}, \mathcal{Q}) = \inf_T \sup_{\theta \in \Theta} \mathsf{TV}(T\mathbb{P}_\theta, \mathbb{Q}_\theta)$, where the infimum is over all Markov kernels, and $T\mathbb{P}_\theta$ denotes the image measure. The Le Cam distance is then $\Delta(\mathcal{P}, \mathcal{Q}) = \delta(\mathcal{P}, \mathcal{Q}) \vee \delta(\mathcal{Q}, \mathcal{P})$. We need the following lemma to prove Lemma 5.1.

Lemma 8.5. *For $X \sim N_p(\mu, \Sigma)$ with $M^{-1} \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq M$ and $U = (U_1, \ldots, U_p)'$ where $U_i \overset{iid}{\sim} \mathrm{Unif}[0, 1]$, we have for any $t^{-1/2}2^t \geq 2(pM)^{3/2}$,*

$$\mathsf{TV}(X, [X]_t + 2^{-t}U) \leq (pM)^{3/2}t^{1/2}2^{-t}.$$

40

Proof. Let f and g denote the density functions of X and $[X]_t + 2^{-t}U$, respectively. Then g is a piecewise constant function. For any $(x_1, ..., x_p) \in B = \prod_{i=1}^{p}[2^{-t}i_j, 2^{-t}(i_j + 1))$, where $i_j \in \mathbb{Z}$, we have

$$g(x_1, ..., x_p) = \frac{1}{\nu(B)} \int_B f(x_1, ..., x_p) dx_1 ... dx_p,$$

where ν is the Lebesgue measure. Hence,

$$\sup_{||x-\mu||_\infty \leq K} \left| \frac{g(x)}{f(x)} - 1 \right| \leq \sup_{\substack{||x-\mu||_\infty \leq K \\ ||x-y||_\infty \leq 2^{-t}}} \left| \frac{f(x)}{f(y)} - 1 \right|$$

$$\leq \sup_{\substack{||x-\mu||_\infty \leq K \\ ||x-y||_\infty \leq 2^{-t}}} \left| e^{|(x-\mu)'\Sigma^{-1}(x-\mu) - (y-\mu)'\Sigma^{-1}(y-\mu)|/2} - 1 \right|$$

$$\leq \sup_{\substack{||x-\mu||_\infty \leq K \\ ||x-y||_\infty \leq 2^{-t}}} \left| e^{||\Sigma^{-1}||_F ||(x-\mu)(x-\mu)' - (y-\mu)(y-\mu)'||_F/2} - 1 \right| \tag{85}$$

$$\leq \exp\left(p^{3/2}MK2^{-t}\right) - 1 \tag{86}$$

$$\leq \frac{3}{2}p^{3/2}MK2^{-t},$$

whenever $p^{3/2}MK2^{-t} \leq \frac{1}{2}$. The inequality (85) holds since

$$|(x-\mu)'\Sigma^{-1}(x-\mu) - (y-\mu)'\Sigma^{-1}(y-\mu)|$$
$$= \mathsf{Tr}\left(\Sigma^{-1}[(x-\mu)(x-\mu)' - (y-\mu)(y-\mu)']\right)$$
$$\leq ||\Sigma^{-1}||_F ||(x-\mu)(x-\mu)' - (y-\mu)(y-\mu)'||_F$$

by Cauchy-Schwarz inequality. The inequality (86) holds because $||\Sigma^{-1}||_F \leq \sqrt{p}||\Sigma^{-1}||_{op} \leq \sqrt{p}M$ and $||(x-\mu)(x-\mu)' - (y-\mu)(y-\mu)'||_F \leq p||x-y||_\infty(||x-\mu||_\infty + ||y-\mu||_\infty) \leq 2pK2^{-t}$. Note that

$$\int |f - g| \leq \int_{||x-\mu||_\infty > K} |f(x) - g(x)|dx + \int_{||x-\mu||_\infty \leq K} f(x) \left| \frac{g(x)}{f(x)} - 1 \right| dx.$$

According to Gaussian tail probability, the first term can be bounded by $2p\sqrt{\frac{2}{\pi}}\frac{\sqrt{M}}{K-1}e^{-\frac{(K-1)^2}{2M}}$. The second term is bounded by $\frac{3}{2}p^{3/2}MK2^{-t}$ according to our previous analysis. Choosing $K = \sqrt{2Mt\log 2} + 1$, we obtain the bound $2(pM)^{3/2}t^{1/2}2^{-t}$ for all $t^{-1/2}2^t \geq 2(pM)^{3/2}$. The conclusion follows the simple fact that $\mathsf{TV}(X, [X]_t + 2^{-t}U) = \frac{1}{2}\int |f - g|$. $\qquad\square$

Proof of Lemma 5.1. Since each distribution in $\mathcal{E}_M^{(p,n,t)}$ comes from discretizing a corresponding distribution in $\mathcal{E}_M^{(p,n)}$ on a grid with equal spacing 2^{-t}, we have $\delta(\mathcal{E}_M^{(p,n)}, \mathcal{E}_M^{(p,n,t)}) = 0$. On the other hand, Lemma 8.5 and Lemma 7 of [26] lead to

$$\delta(\mathcal{E}_M^{(p,n,t)}, \mathcal{E}_M^{(p,n)}) \leq n(pM)^{3/2}t^{1/2}2^{-t}.$$

This completes the proof. $\qquad\square$

41

8.3.2 Randomized polynomial-time reduction for discretized models

For the discretized model, we modify the reduction scheme in Sections 5.2 and 7 and turn them into randomized polynomial-time reductions.

Truncated dyadic approximations To this end, we first introduce the following truncated dyadic approximation for any univariate distribution $\mathcal{F}$ with density f. For any $w, K \in \mathbb{N}$ and $K + w + 1 < b \in \mathbb{N}$, define the discrete distribution $\mathcal{A}_{w,b,K}[\mathcal{F}]$ with probability mass function $\mathcal{A}_{w,b,K}[f]$ as

$$\mathcal{A}_{w,b,K}[f](-2^K + (i-1)2^{-w}) = \left[\frac{\int_{-2^K+(i-1)2^{-w}}^{-2^K+i2^{-w}} f(x)\mathrm{d}x}{\int_{-2^K}^{2^K} f(x)\mathrm{d}x} \right]_b, \quad i \in [2^{K+w+1}-1], \qquad (87)$$

and let

$$\mathcal{A}_{w,b,K}[f](2^K - 2^{-w}) = 1 - \sum_{i=1}^{2^{K+w+1}-1} \mathcal{A}_{w,b,K}[f](-2^K + (i-1)2^{-w}). \qquad (88)$$

In (87), $[\cdot]_b$ is the quantization defined previously in (24), and (88) ensures that $\mathcal{A}_{w,b,K}[\mathcal{F}]$ is a proper probability distribution. Albeit the relatively complicated expressions in (87) and (88), the distribution $\mathcal{A}_{w,b,K}[\mathcal{F}]$ can be obtained in the following way. For any random variable $U \sim \mathcal{F}$, we have the distribution of $[U\mathbf{1}_{\{U \in [-2^K, 2^K]\}}]_w$ having probability mass function

$$\mathbb{P}([U\mathbf{1}_{\{U\in[-2^K,2^K]\}}]_w = -2^K + (i-1)2^{-w}) = p_i = \frac{\int_{-2^K+(i-1)2^{-w}}^{-2^K+i2^{-w}} f(x)\mathrm{d}x}{\int_{-2^K}^{2^K} f(x)\mathrm{d}x}$$

for $i = 1, \ldots, 2^{K+w+1}$. Then (87) and (88) are obtained by replacing $(p_1, \ldots, p_{2^{K+w+1}})$ with its dyadic approximation

$$([p_1]_b, \ldots, [p_{2^{K+w+1}-1}]_b, 1 - \sum_{i=1}^{2^{K+w+1}-1}[p_i]_b). \qquad (89)$$

Remark 8.1. By the definition of total variation distance, it is straightforward to verify that the approximation error in total variation distance by (89) is upper bounded by $2^{K+w+1-b}$. As discussed in Section 4.2 of [26], regardless of the original distribution $\mathcal{F}$, the computational complexity of drawing a random number from $\mathcal{A}_{w,b,K}(\mathcal{F})$ is $O(b2^{K+w})$. This fact is crucial in ensuring the modified reduction below is of randomized polynomial-time.

Randomized polynomial-time reduction Let $t = \lceil 4\log_2(p + m + n) \rceil$ in the case of CCA (and $t = \lceil 4\log_2(p+n) \rceil$ in the case of PCA),

$$\begin{aligned} w &= t + \lceil 4\log_2 p \rceil, \; K = \lceil \log_2(3\sqrt{\log(N+p)}) \rceil, \\ b &= w + K + 1 + \lceil 4\log_2 p \rceil. \end{aligned} \qquad (90)$$

With the above choice of w, b and K, in the case of sparse CCA, we apply the following modifications to the four steps:

1. Initialization. We sample i.i.d. r.v.'s $\check{\xi}_1, \ldots, \check{\xi}_{2n} \sim \mathcal{A}_{w,b,K}[\widetilde{\Phi}_0]$ and set $\check{\mu}_i = [\eta_N^{1/2}]_w \check{\xi}_i$ for $i \in [2n]$.

2. Gaussianization. We generate two $2n \times 2n$ matrices $\check{B}_0, \check{B}_1$ where conditioning on the $\check{\mu}_i$'s all entries are mutually independent satisfying

$$\mathcal{L}((\check{B}_0)_{ij}|\check{\mu}_i) = \mathcal{A}_{w,b,K}[\mathcal{F}_{\mu_i,0}], \quad \mathcal{L}((\check{B}_1)_{ij}|\check{\mu}_i) = \mathcal{A}_{w,b,K}[\mathcal{F}_{\mu_i,1}].$$

We then generate a matrix $\check{W}$ of size $2n \times p$ where

$$\check{W}_{ij} = (\check{B}_0)_{ij}\,(1 - (A_0)_{ij}) + (\check{B}_1)_{ij}(A_0)_{ij}, \quad \text{for } i \in [2n],\ j \in [2n].$$

When $2n < j \leq p$, we let $\check{W}_{ij}$ be independent draws from $\mathcal{A}_{w,b,K}[N(0,1)]$.

3. Sample Generation. Let $\check{W}_i$ be the i-th row of $\check{W}$. For $i \in [n]$, we generate independent random vector $\check{Z}_i = (\check{Z}_{i1}, \ldots, \check{Z}_{ip})$ where $\check{Z}_{ij} \overset{iid}{\sim} \mathcal{A}_{w,b,K}[N(0,1)]$. Define

$$\check{X}_i = \frac{1}{[\sqrt{2}]_w}(\check{W}_{n+i} + \check{Z}_i), \quad \check{Y}_i = \frac{1}{[\sqrt{2}]_w}(\check{W}_{n+i} - \check{Z}_i).$$

Let $\check{X} = [\check{X}_1', \ldots, \check{X}_n']'$ and $\check{Y} = [\check{Y}_1', \ldots, \check{Y}_n']'$.

4. Test Construction. Let $\widehat{u} = \widehat{u}([\check{X}]_t, [\check{Y}]_t)$ by treating $\{([\check{X}_i]_t, [\check{Y}_i]_t)\}_{i=1}^n$ as data. We reject H_0^G if

$$\frac{\widehat{u}'(\frac{1}{n}\sum_{i=1}^n [\check{W}_i]_t [\check{W}_i]_t')\widehat{u}}{\|\widehat{u}\|^2} \geq 1 + \frac{1}{4}k\eta_N.$$

In the case of sparse PCA, the first two steps of the reduction are the same as above. In the third step, denote the i-th row of $\check{W}$ by $\check{W}_i$. Let $\widehat{\theta} = \widehat{\theta}([\check{W}_{n+1}]_t, \ldots, [\check{W}_{2n}]_t)$ be the estimator of θ by treating $\{[\check{W}_{n+i}]_t\}_{i=1}^n$ as data. We reject H_0^G if

$$\frac{\widehat{\theta}'(\frac{1}{n}\sum_{i=1}^n [\check{W}_i]_t [\check{W}_i]_t')\widehat{\theta}}{\|\widehat{\theta}\|^2} \geq 1 + \frac{1}{4}k\eta_N.$$

Remark 8.2. We now verify that under the conditions of Theorems 5.1 and 7.1 and the choice of w, b and K in (90), the modified reductions stated above are of randomized polynomial time. First, by Remark 8.1 and (90), the complexity for sampling any random variable in the above reduction is $O(p^8(\log p)^{3/2})$, and in total, we need to generate no more than $O(n(p+n))$ random variables. Hence, the total complexity for random number generation is $O(p^{10}(\log p)^{3/2})$ in view of the condition $p \geq 2n$. On the other hand, it is straightforward to verify that all the other computations (except for the estimator $\widehat{u}$ or $\widehat{\theta}$) have complexity no more than $O(p^{10}(\log p)^{3/2})$. Since the conditions of Theorems 5.1 and 7.1 ensure that for some constant $a > 1$, $2n \leq p \leq n^a$ and $n \leq N/12$, we obtain that the additional computational complexity induced by the proposed reductions is $O(N^{10a}(\log N)^{3/2})$. Therefore, they are of randomized polynomial-time.

Intermediate results for discrete data We now present results for discrete data which are in parallel to those in Lemmas 8.3 and 8.4 and Theorems 5.2 and 7.2 for continuous data. The proofs of these results can be modified from those of Lemmas 8.3 and 8.4 and Theorems 5.2 and 7.2 in essentially the way as was did in turning the proofs of Lemmas 3-4 and Theorem 3 to those of Lemmas 5-6 and Theorem 4 in [26], and hence are omitted.

To state the results, let $\mathbb{P}_0^{t,n}$, $\mathbb{Q}_0^{t,n}$, $\mathbb{P}_{\theta,\tau}^{t,n}$ and $\mathbb{P}_{\theta,\tau}^{t,n}$ be the discretized versions of $\mathbb{P}_0^n$, $\mathbb{Q}_0^n$, $\mathbb{P}_{\theta,\tau}^n$ and $\mathbb{P}_{\theta,\tau}^n$. The following two lemmas are in parallel to Lemmas 8.3 and 8.4.

Lemma 8.6. *Suppose $A \sim \mathcal{G}(N, 1/2)$. Then there exists some constant $C_1 > 0$, such that*

$$\mathsf{TV}(\mathcal{L}(\{[\check{W}_i]_t\}_{i=1}^{2n}), \mathbb{Q}_0^{t,2n}) \leq C_1 N^{-1},$$
$$\mathsf{TV}(\mathcal{L}(\{[\check{X}_i]_t, [\check{Y}_i]_t\}_{i=1}^n, \{[\check{W}_i]_t\}_{i=1}^n), \mathbb{P}_0^{t,n} \times \mathbb{Q}_0^{t,n}) \leq C_1 N^{-1}.$$

Lemma 8.7. *Suppose $A \sim \mathcal{G}(N, 1/2, k)$. Then there exists a distribution π supported on the set (76) such that for some absolute constants $C_1, C_2 > 0$,*

$$\mathsf{TV}(\mathcal{L}(\{[\check{W}_i]_t\}_{i=1}^{2n}), \int \mathbb{Q}_{\theta,\tau}^{t,2n} \mathrm{d}\pi(\theta,\tau)) \leq C_1 \left(e^{-C_2 k} + \frac{1}{N}\right) + \frac{4n}{N},$$

$$\mathsf{TV}(\mathcal{L}(\{([\check{X}_i]_t, [\check{Y}_i]_t)\}_{i=1}^n, \{[\check{W}_i]_t\}_{i=1}^n), \int (\mathbb{P}_{\theta,\tau}^{t,n} \times \mathbb{Q}_{\theta,\tau}^{t,n}) \mathrm{d}\pi(\theta,\tau))$$
$$\leq C_1 \left(e^{-C_2 k} + \frac{1}{N}\right) + \frac{4n}{N}.$$

The next two theorems are in parallel to Theorems 5.2 and 7.2, respectively.

Theorem 8.1. *Let k, N, n, p satisfy the condition of Theorem 5.2. For any randomized polynomial-time estimator $\widehat{u}$ satisfying*

$$\sup_{\mathbb{P} \in \mathcal{P}^t(n, 3k/2, 3k/2, p, p, 1, k\eta_N/8; 4)} \mathbb{P}\left\{L(\widehat{u}, u) > \frac{1}{3 \times 32^2}\right\} \leq \beta, \tag{91}$$

for $t = \lceil 4\log_2(p+m+n)\rceil$, there exists a randomized polynomial-time test ψ for (22) satisfying

$$\mathbb{P}_{H_0^G}\psi + \mathbb{P}_{H_1^G}(1 - \psi) < \beta + \frac{4n}{N} + C(n^{-1} + N^{-1} + e^{-C'k}),$$

for sufficiently large n with some constants $C, C' > 0$.

Theorem 8.2. *Let k, N, n, p satisfy the condition of Theorem 7.2. For any randomized polynomial-time estimator $\widehat{\theta}$ satisfying*

$$\sup_{\mathbb{Q} \in \mathcal{Q}^t(n, 3k/2, p, 1, k\eta_N/2; 3)} \mathbb{Q}\left\{\|P_{\widehat{\theta}} - P_\theta\|_F^2 > \frac{1}{3}\right\} \leq \beta,$$

for $t = \lceil 4\log_2(p+n)\rceil$, there exists a randomized polynomial-time test ψ for (22) satisfying

$$\mathbb{P}_{H_0^G}\psi + \mathbb{P}_{H_1^G}(1 - \psi) < \beta + \frac{4n}{N} + C(n^{-1} + N^{-1} + e^{-C'k}),$$

for sufficiently large n with some constants $C, C' > 0$.

8.3.3 Proofs of Theorems 5.1 and 7.1

We present below the proof of Theorem 5.1 and Theorem 7.1 can be proved in a similar way.

Proof of Theorem 5.1. Suppose there existed a randomized polynomial-time estimator $\widehat{u}$ such that as $n \to \infty$, (25) holds and

$$\liminf_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}^t(n,s_u,s_v,p,m,1,\lambda;4)} \mathbb{P}\left\{L(\widehat{u},u) > \frac{1}{3\times 32^2}\right\} \leq \frac{1}{4}. \tag{92}$$

Now let $N = 12n$ and $k = \lfloor 2s_u/3 \rfloor$. Then Theorem 8.1 implies that there exists a randomized polynomial-time test ψ for (22) such that

$$\liminf_{n\to\infty}\left(\mathbb{P}_{H_0^G}\psi + \mathbb{P}_{H_1^G}(1-\psi)\right) \leq \frac{1}{4} + \frac{4}{12} < \frac{2}{3}. \tag{93}$$

On the other hand, (25) and the conditions of the theorem implies

$$\limsup_{n\to\infty} \frac{\log k}{\log N} < \frac{1}{2}.$$

This contradicts Hypothesis A and hence completes the proof. □

9 Additional Proofs

9.1 Proof of Theorem 3.1

In this section, we denote $\widehat{\Sigma}_x^{(0)}$, $\widehat{\Sigma}_y^{(0)}$ and $\widehat{\Sigma}_{xy}^{(0)}$ by $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$ and $\widehat{\Sigma}_{xy}$ for simplicity of notation.

Lemma 9.1. *Assume* $\frac{1}{n}\left(s_u\log(ep/s_u) + s_v\log(em/s_v)\right) \leq c$ *for some sufficiently small* $c > 0$. *Then, for any* $C' > 0$, *there exists* $C > 0$ *only depending on* C' *such that*

$$\begin{aligned}(1-\delta_C)\|\Sigma_x^{1/2}(\widehat{U}\widehat{V}' - \widetilde{U}\widetilde{V}')\Sigma_y^{1/2}\|_F^2 &\leq& \|\widehat{\Sigma}_x^{1/2}(\widehat{U}\widehat{V}' - \widetilde{U}\widetilde{V}')\widehat{\Sigma}_y^{1/2}\|_F^2 \\ &\leq& (1+\delta_C)\|\Sigma_x^{1/2}(\widehat{U}\widehat{V}' - \widetilde{U}\widetilde{V}')\Sigma_y^{1/2}\|_F^2,\end{aligned}$$

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

$$\delta_C = C\left[\sqrt{\frac{s_u\log(ep/s_u)}{n}} + \sqrt{\frac{s_v\log(em/s_v)}{n}}\right].$$

For the following two lemmas, we use the notation

$$\epsilon_n^2 = \frac{1}{n}\left(r(s_u + s_v) + s_u\log\frac{ep}{s_u} + s_v\log\frac{em}{s_v}\right).$$

The following two lemmas are slight variations of Lemma 5 and Lemma 6 in [16], and thus we omit their proofs.

Lemma 9.2. *Assume $\frac{1}{n}\left(r(s_u + s_v) + s_u\log(ep/s_u) + s_v\log(em/s_v)\right) \le c$ for some sufficiently small $c > 0$. Then, for any $C' > 0$, there exists $C > 0$ only depending on C' such that*

$$\left|\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, \widetilde{U}\widetilde{V}' - \widehat{U}\widehat{V}' \rangle\right| \le C\epsilon_n \|\Sigma_x^{1/2}(\widehat{U}\widehat{V}' - \widetilde{U}\widetilde{V}')\Sigma_y^{1/2}\|_F,$$

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

Lemma 9.3. *Assume $\frac{1}{n}\left(r(s_u + s_v) + s_u\log(ep/s_u) + s_v\log(em/s_v)\right) \le c$ for some sufficiently small $c > 0$. Then, for any $C' > 0$, there exists $C > 0$ only depending on C' such that*

$$\left|\langle \widehat{\Sigma}_x U\Lambda V'\widehat{\Sigma}_y - \Sigma_x U\Lambda V'\Sigma_y, \widetilde{U}\widetilde{V}' - \widehat{U}\widehat{V}' \rangle\right| \le C\epsilon_n \|\Sigma_x^{1/2}(\widehat{U}\widehat{V}' - \widetilde{U}\widetilde{V}')\Sigma_y^{1/2}\|_F,$$

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

Proof of Theorem 3.1. We use the notation $\Delta = U^{(0)}(\widehat{V}^{(0)})' - \widetilde{U}\widetilde{V}'$. Let us first derive a bound for $\|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_F^2$. We have

$$
\begin{aligned}
&\|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_F^2 \\
\le\ & 2\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F^2 && (94) \\
\le\ & \frac{4}{\lambda_r}\langle \widehat{\Sigma}_x \widetilde{U}\widetilde{\Lambda}\widetilde{V}'\widehat{\Sigma}_y, -\Delta\rangle + \frac{4}{\lambda_r}\|\widetilde{\Lambda} - \Lambda\|_F\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F && (95) \\
=\ & \frac{4}{\lambda_r}\langle \widehat{\Sigma}_x U\Lambda V'\widehat{\Sigma}_y, -\Delta\rangle + \frac{4}{\lambda_r}\|\widetilde{\Lambda} - \Lambda\|_F\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F \\
\le\ & \frac{4}{\lambda_r}\langle \widehat{\Sigma}_x U\Lambda V'\widehat{\Sigma}_y - \widehat{\Sigma}_{xy}, -\Delta\rangle + \frac{4}{\lambda_r}\|\widetilde{\Lambda} - \Lambda\|_F\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_F && (96) \\
\le\ & \frac{4}{\lambda_r}\langle \widehat{\Sigma}_x U\Lambda V'\widehat{\Sigma}_y - \Sigma_x U\Lambda V'\Sigma_y, -\Delta\rangle \\
& + \frac{4}{\lambda_r}\langle \Sigma_{xy} - \widehat{\Sigma}_{xy}, -\Delta\rangle + \frac{8}{\lambda_r}\|\widetilde{\Lambda} - \Lambda\|_F\|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_F.
\end{aligned}
$$

In the above argument, we have used Lemma 9.1 to obtain (94). The inequality (95) is due to Lemma 6.3, and the inequality (96) is because of the fact that

$$\langle \widehat{\Sigma}_{xy}, -\Delta\rangle \le 0,$$

by the definition of the estimator. Let us use the notation $L = \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_F$ and $\epsilon_n^2 = \frac{1}{n}\left(r(s_u + s_v) + s_u\log\frac{ep}{s_u} + s_v\log\frac{em}{s_v}\right)$. By Lemma 6.1, Lemma 9.2 and Lemma 9.3, we have

$$L^2 \le \frac{16C\epsilon_n L}{\lambda_r},$$

with high probability. This leads to $L^2 \le C_1(\epsilon_n/\lambda)^2$ with high probability. By triangle inequality, we have

$$\|\Sigma_x^{1/2}(U^{(0)}(\widehat{V}^{(0)})' - UV')\Sigma_y^{1/2}\|_F \le \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_F + \|\Sigma_x^{1/2}(\widetilde{U}\widetilde{V}' - UV')\Sigma_y^{1/2}\|_F.$$

Using Lemma 6.1 and $L^2 \le C_1(\epsilon_n/\lambda)^2$, we complete the proof. $\qquad\square$

9.2 Proof of Theorem 3.2

In this section, we denote $\widehat{\Sigma}_x^{(1)}$, $\widehat{\Sigma}_y^{(1)}$ and $\widehat{\Sigma}_{xy}^{(1)}$ by $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$ and $\widehat{\Sigma}_{xy}$ for simplicity of notation. Let $U^* = U\Lambda V' \Sigma_y \widehat{V}^{(0)}$, and $\Delta = \widehat{U}^{(1)} - U^*$.

Lemma 9.4. *Assume* $\frac{s_u \log(ep/s_u)}{n} \leq c$ *for some sufficiently small* $c > 0$. *Then, for any* $C' > 0$, *there is some* $C > 0$ *only depending on* C', *such that*

$$\|\Sigma_y^{1/2} \widehat{V}^{(0)}\|_{\mathrm{op}} \leq 1 + C\sqrt{\frac{s_u \log(ep/s_u)}{n}},$$

$$\|(\widehat{V}^{(0)})'\Sigma_y \widehat{V}^{(0)} - I\|_{\mathrm{op}} \leq C\sqrt{\frac{s_u \log(ep/s_u)}{n}},$$

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

Lemma 9.5. *Assume* $\frac{s_u \log(ep/s_u)}{n} \leq c$ *for some sufficiently small* $c > 0$. *Then, for any* $C' > 0$, *there is some* $C > 0$ *only depending on* C', *such that*

$$(1 - \delta'_C)\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}}^2 \leq \|\widehat{\Sigma}_x^{1/2}\Delta\|_{\mathrm{F}}^2 \leq (1 + \delta'_C)\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}}^2,$$

with probability at least $1 - \exp(-C's_u \log(ep/s_u))$, *with* $\delta'_C = C\sqrt{\frac{s_u \log(ep/s_u)}{n}}$.

Lemma 9.6. *Assume* $\frac{1}{n}(s_v \log(em/s_v) + s_u \log(ep/s_u) + rs_u) \leq c$ *for some sufficiently small* $c > 0$. *Then, for any* $C' > 0$, *there is some* $C > 0$ *only depending on* C', *such that*

$$\left|\mathrm{Tr}\left(\Delta'(\widehat{\Sigma}_{xy} - \Sigma_{xy})\widehat{V}^{(0)}\right)\right| \leq C\sqrt{\frac{rs_u + s_u \log(ep/s_u)}{n}}\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}},$$

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

Lemma 9.7. *Assume* $\frac{1}{n}(s_v \log(em/s_v) + s_u \log(ep/s_u) + rs_u) \leq c$ *for some sufficiently small* $c > 0$. *Then for any* $C' > 0$, *there is some* $C > 0$ *only depending on* C', *such that*

$$\left|\mathrm{Tr}\left(\Delta'(\widehat{\Sigma}_x - \Sigma_x)U^*\right)\right| \leq C\sqrt{\frac{rs_u + s_u \log(ep/s_u)}{n}}\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}},$$

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

Proof. The proof consists of two steps. In the first step, we derive a bound for $\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}}$. In the second step, we derive the desired bound for $L(\widehat{U}, U)$.

Step 1. By the definition of the estimator, we have

$$\mathrm{Tr}((\widehat{U}^{(1)})'\widehat{\Sigma}_x\widehat{U}^{(1)}) - 2\,\mathrm{Tr}((\widehat{U}^{(1)})'\widehat{\Sigma}_{xy}\widehat{V}^{(0)}) \leq \mathrm{Tr}((U^*)'\widehat{\Sigma}_x U^*) - 2\,\mathrm{Tr}((U^*)'\widehat{\Sigma}_{xy}\widehat{V}^{(0)}).$$

After rearrangement, we have

$$\begin{aligned}
&\mathrm{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \\
&\leq\ 2\,\mathrm{Tr}\left(\Delta'(\widehat{\Sigma}_{xy}\widehat{V}^{(0)} - \widehat{\Sigma}_x U^*)\right) \\
&\leq\ 2\left|\mathrm{Tr}\left(\Delta'(\widehat{\Sigma}_{xy} - \Sigma_{xy})\widehat{V}^{(0)}\right)\right| + 2\left|\mathrm{Tr}\left(\Delta'(\widehat{\Sigma}_x - \Sigma_x)U^*\right)\right|.
\end{aligned}$$

Using Lemma 9.5, Lemma 9.6 and Lemma 9.7, we have

$$\frac{1}{2}\|\Sigma_x\Delta\|_{\mathrm{F}}^2 \leq 4C\sqrt{\frac{rs_u + s_u\log(ep/s_u)}{n}}\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}},$$

with high probability, which immediately implies a bound for $\|\Sigma_x\Delta\|_{\mathrm{F}}^2$. This completes Step 1.

Step 2. We claim that

$$\sigma_{\min}^{-1}\left(\Sigma_x^{1/2}U\Lambda V'\Sigma_y\widehat{V}^{(0)}\right) \leq \frac{C}{\lambda}, \tag{97}$$

$$\|\Sigma_x^{1/2}\widehat{U} - \Sigma_x^{1/2}\widehat{U}^{(1)}((\widehat{U}^{(1)})'\Sigma_x\widehat{U}^{(1)})^{-1/2}\|_{\mathrm{F}} \leq C\sqrt{\frac{rs_u + s_u\log(ep/s_u)}{n}}, \tag{98}$$

with high probability. The two claims (97) and (98) will be proved in the end. We bound $L(\widehat{U},U)$ by

$$\begin{aligned}
\sqrt{L(\widehat{U},U)} &= \inf_{W\in O(r,r)}\|\Sigma_x^{1/2}(\widehat{U}W - U)\|_{\mathrm{F}} \\
&\leq \|\Sigma_x^{1/2}\widehat{U} - \Sigma_x^{1/2}\widehat{U}^{(1)}((\widehat{U}^{(1)})'\Sigma_x\widehat{U}^{(1)})^{-1/2}\|_{\mathrm{F}} \\
&\quad + \inf_{W\in O(r,r)}\|\Sigma_x^{1/2}\widehat{U}^{(1)}((\widehat{U}^{(1)})'\Sigma_x\widehat{U}^{(1)})^{-1/2}W - \Sigma_x^{1/2}U\|_{\mathrm{F}} \\
&\leq C\sqrt{\frac{rs_u + s_u\log(ep/s_u)}{n}} + \frac{1}{\sqrt{2}}\|P_{\Sigma_x^{1/2}\widehat{U}} - P_{\Sigma_x^{1/2}U}\|_{\mathrm{F}} \tag{99} \\
&\leq C\sqrt{\frac{rs_u + s_u\log(ep/s_u)}{n}} + C\sigma_{\min}^{-1}\left(\Sigma_x^{1/2}U\Lambda V'\Sigma_y\widehat{V}^{(0)}\right)\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}} \tag{100} \\
&\leq C\sqrt{\frac{rs_u + s_u\log(ep/s_u)}{n\lambda^2}},
\end{aligned}$$

with high probability. The inequality (99) is due to the claim (98), Lemma 6.6 and the fact that $P_{\Sigma^{1/2}\widehat{U}} = P_{\Sigma^{1/2}\widehat{U}^{(1)}}$. The inequality (100) is derived from the sin-theta theorem [36]. Thus, we have obtained the desired bound for the loss $L(\widehat{U},U)$. To finish the proof, we need to prove the two claims (97) and (98). Since $\Sigma_x^{1/2}U \in O(p,r)$, we have

$$\sigma_{\min}^{-1}\left(\Sigma_x^{1/2}U\Lambda V'\Sigma_y\widehat{V}^{(0)}\right) \leq \lambda^{-1}\|(V'\Sigma_y\widehat{V}^{(0)})^{-1}\|_{\mathrm{op}}.$$

Thus, it is sufficient to bound $\|(V'\Sigma_y\widehat{V}^{(0)})^{-1}\|_{\mathrm{op}}$. By Theorem 3.1 and sin-theta theorem [36], $\|P_{\Sigma_y^{1/2}\widehat{V}^{(0)}} - P_{\Sigma_y^{1/2}V}\|_{\mathrm{F}}$ is sufficiently small. In view of Lemma 6.6, there exists $W \in O(r,r)$, such that

$$\|\Sigma_y^{1/2}\widehat{V}^{(0)}(\widehat{V}^{(0)}\Sigma_y\widehat{V}^{(0)})^{-1/2} - \Sigma_y^{1/2}VW\|_{\mathrm{F}}$$

is sufficiently small. Therefore, together with Lemma 9.4,

$$\begin{aligned}
&\|V'\Sigma_y\widehat{V}^{(0)} - W\|_{\mathrm{op}} \\
&\leq \|V'\Sigma_y\widehat{V}^{(0)} - V'\Sigma_yVW(\widehat{V}^{(0)}\Sigma_y\widehat{V}^{(0)})^{1/2}\|_{\mathrm{op}} + \|W\|_{\mathrm{op}}\|(\widehat{V}^{(0)}\Sigma_y\widehat{V}^{(0)})^{1/2} - I\|_{\mathrm{op}} \\
&\leq \|\Sigma_y^{1/2}\widehat{V}^{(0)}(\widehat{V}^{(0)}\Sigma_y\widehat{V}^{(0)})^{-1/2} - \Sigma_y^{1/2}VW\|_{\mathrm{F}}\|(\widehat{V}^{(0)}\Sigma_y\widehat{V}^{(0)})^{1/2}\|_{\mathrm{op}} + \|(\widehat{V}^{(0)}\Sigma_y\widehat{V}^{(0)})^{1/2} - I\|_{\mathrm{op}}
\end{aligned}$$

48

is also sufficiently small. By Weyl's inequality [17, p.449], $|\sigma_{\min}(V'\Sigma_y\widehat{V}^{(0)}) - 1| \leq \|V'\Sigma_y\widehat{V}^{(0)} - W\|_{\mathrm{op}}$ is sufficiently small. Hence, $\|(V'\Sigma_y\widehat{V}^{(0)})^{-1}\|_{\mathrm{op}} \leq 2$ with high probability, which implies the desired bound in (97). Finally, we need to prove (98). We have

$$
\|\Sigma_x^{1/2}\widehat{U} - \Sigma_x^{1/2}\widehat{U}^{(1)}((\widehat{U}^{(1)})'\Sigma_x\widehat{U}^{(1)})^{-1/2}\|_{\mathrm{F}}
$$
$$
\leq \quad \|\Sigma_x^{1/2}\widehat{U}^{(1)}\|_{\mathrm{F}}\|((\widehat{U}^{(1)})'\widehat{\Sigma}_x^{(2)}\widehat{U}^{(1)})^{-1/2} - ((\widehat{U}^{(1)})'\Sigma_x\widehat{U}^{(1)})^{-1/2}\|_{\mathrm{op}}
$$
$$
\leq \quad C\left(\|\Sigma_x^{1/2}U\Lambda V'\Sigma_y\widehat{V}^{(0)}\|_{\mathrm{F}} + \|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}}\right)\|(\widehat{U}^{(1)})'(\widehat{\Sigma}_x^{(2)} - \Sigma_x)\widehat{U}^{(1)}\|_{\mathrm{op}}.
$$

We have already shown that $\|\Sigma_x^{1/2}\Delta\|_{\mathrm{F}}$ is sufficiently small. The term $\|\Sigma_x^{1/2}U\Lambda V'\Sigma_y\widehat{V}^{(0)}\|_{\mathrm{F}}$ is bounded by $\sqrt{r}\|V'\Sigma_y\widehat{V}^{(0)}\|_{\mathrm{op}} \leq \sqrt{r}(1 + \|V'\Sigma_y\widehat{V}^{(0)} - W\|_{\mathrm{op}}) \leq C\sqrt{r}$ by using the bound derived for $\|V'\Sigma_y\widehat{V}^{(0)} - W\|_{\mathrm{op}}$. To bound $\|(\widehat{U}^{(1)})'(\widehat{\Sigma}_x^{(2)} - \Sigma_x)\widehat{U}^{(1)}\|_{\mathrm{op}}$, note that $\widehat{\Sigma}_x^{(2)}$ only depends on $\mathcal{D}_2$ and is independent of $\widehat{U}^{(1)}$. Using union bound and an ϵ-net argument (see, for example, [30]) and the fact that $r \leq s_u$ (which is implied by $\Sigma_x^{1/2}U \in O(p,r)$), we have $\|(\widehat{U}^{(1)})'(\widehat{\Sigma}_x^{(2)} - \Sigma_x)\widehat{U}^{(1)}\|_{\mathrm{op}} \leq C\sqrt{\frac{rs_u + s_u \log(ep/s_u)}{n}}$ with high probability. Hence, the proof is complete. $\qquad\square$

9.3 Proof of Theorem 3.3

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 14 in [16]. It gives explicit formula for the Kullback-Leibler divergence between distributions generated by a special kind of covariance matrices.

Lemma 9.8. *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*

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

The main tool for our proof is Fano's lemma. The following version is adapted from [39, Lemma 3].

Proposition 9.1. *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*

$$
d_{\mathrm{KL}}(T) \triangleq \sup_{\theta, \theta' \in T} D(\mathbb{P}_\theta\|\mathbb{P}_{\theta'}). \tag{101}
$$

Then

$$
\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(\rho^2\big(\widehat{\theta}(X), \theta\big) \geq \frac{\epsilon^2}{4}\right) \geq 1 - \frac{d_{\mathrm{KL}}(T) + \log 2}{\log \mathcal{M}(T, \rho, \epsilon)}. \tag{102}
$$

Finally, we lower bound the prediction loss by the squared subspace distance. Its proof is given in Section 9.4.

Proposition 9.2. *There exists a constant $C > 0$ only depending on M, such that*

$$\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \le CL(\widehat{U}, U).$$

A similar inequality holds for $L(\widehat{V}, V)$.

Proof of Theorem 3.3. Let us first give an outline of the proof. By Proposition 9.2, we have

$$\inf_{\widehat{U}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\left(L(\widehat{U}, U) \ge C\epsilon^2\right) \ge \inf_{\widehat{U}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \ge C_1\epsilon^2\right),$$

for any rate ϵ^2. Therefore, it is sufficient to derive a lower bound for the loss $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$. Without loss of generality, we assume $s_u/3$ is an integer and $s_u \le 3p/4$. The case $s_u > 3p/4$ is harder and thus it shares the same lower bound. The subset of covariance class $\mathcal{F}(p, m, s_u, s_v, r, \lambda; M)$ we consider is

$$T = \left\{ \Sigma = \begin{bmatrix} I_p & \lambda U V_0' \\ \lambda V_0 U' & I_m \end{bmatrix} : U = \begin{bmatrix} \widetilde{U} & 0 \\ 0 & u_r \end{bmatrix}, \widetilde{U} \in B, \right.$$

$$\left. u_r \in \mathbb{R}^{p - 2s_u/3}, \|u_r\| = 1, |\mathrm{supp}(u_r)| \le s_u/3 \right\},$$

where $V_0 = \begin{bmatrix} I_r \\ 0 \end{bmatrix} \in O(m, r)$ and B is a subset of $O(2s_u/3, r-1)$ to be specified later. From the construction, U depends on the matrix $\widetilde{U}$ and the vector u_r. As $\widetilde{U}$ and u_r vary, we always have $U \in O(p, r)$. We use $T(u_r^*)$ to denote a subset of T where $u_r = u_r^*$ is fixed, and use $T(\widetilde{U}^*)$ to denote a subset of T where $\widetilde{U} = \widetilde{U}^*$ is fixed.

The proof has three steps. In the first step, we derive the part $\frac{r s_u}{n\lambda^2}$ using the subset $T(u_r^*)$ for some particular u_r^*. In the second step, we derive the other part $\frac{s_u \log(ep/s_u)}{n\lambda^2}$ using the subset $T(\widetilde{U}^*)$ for some fixed $\widetilde{U}^*$. Finally, we combine the two results in the third step.

Step 1. Let $u_r^* = (1, 0, ..., 0)'$, and we consider the subset $T(u_r^*)$. Let $\widetilde{U}_0 = \begin{bmatrix} I_{r-1} \\ 0 \end{bmatrix} \in O(2s_u/3, r-1)$ and $\epsilon_0 \in (0, \sqrt{r}]$ to be specified later. Define

$$B = B(\epsilon_0) = \left\{ \widetilde{U} \in O(2s_u/3, r-1) : \|\widetilde{U} - \widetilde{U}_0\|_{\mathrm{F}} \le \epsilon_0 \right\}.$$

By Lemma 9.8,

$$d_{\mathrm{KL}}\left(T(u_r^*)\right) = \sup_{\widetilde{U}_{(i)} \in B(\epsilon_0)} \frac{n\lambda^2}{2(1 - \lambda^2)} \|\widetilde{U}_{(1)} - \widetilde{U}_{(2)}\|_{\mathrm{F}}^2 \le \frac{2n\lambda^2 \epsilon_0^2}{1 - \lambda^2}. \tag{103}$$

Here, the equality is due to the definition of V_0 and the inequality due to the definition of $B(\epsilon_0)$. We now establish a lower bound for the packing number of $T(u_r^*)$. For some $\alpha \in (0, 1)$

50

to be specified later, let $\{\widetilde{U}_{(1)}, \ldots, \widetilde{U}_{(N)}\} \subset O(2s_u/3, r-1)$ be a maximal set such that for any $i \neq j \in [N]$,

$$\|\widetilde{U}_{(i)}\widetilde{U}'_{(i)} - \widetilde{U}_0\widetilde{U}'_0\|_{\mathrm{F}} \leq \epsilon_0, \qquad \|\widetilde{U}_{(i)}\widetilde{U}'_{(i)} - \widetilde{U}_{(j)}\widetilde{U}'_{(j)}\|_{\mathrm{F}} \geq \sqrt{2}\alpha\epsilon_0. \tag{104}$$

Then by [11, Lemma 1], for some absolute constant $C > 1$,

$$N \geq \left(\frac{1}{C\alpha}\right)^{(r-1)(2s_u/3-r+1)}.$$

It is easy to see that the loss function $\|P_{U_{(i)}} - P_{U_{(j)}}\|_{\mathrm{F}}^2$ on the subset $T(u_r^*)$ equals $\|\widetilde{U}_{(i)}\widetilde{U}'_{(i)} - \widetilde{U}_{(j)}\widetilde{U}'_{(j)}\|_{\mathrm{F}}^2$. Thus, for $\epsilon = \sqrt{2}\alpha\epsilon_0$ with sufficiently small α, $\log \mathcal{M}(T(u_r^*), \rho, \epsilon) \geq (r-1)(2s_u/3 - r+1)\log\frac{1}{C\alpha} \geq (r-1)(\frac{1}{6}s_u - 1)\log\frac{1}{C\alpha} \geq \frac{1}{12}rs_u \log\frac{1}{C\alpha}$ when r is sufficiently large and $r \leq s_u/2$. Taking $\epsilon_0^2 = c_1\frac{rs_u}{n\lambda^2}$ for sufficiently small c_1, we have

$$\inf_{\widehat{U}} \sup_{T(u_r^*)} \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq \frac{\epsilon^2}{4}\right) \geq 1 - \frac{\frac{2c_1 rs_u}{1-\lambda^2} + \log 2}{\frac{1}{12}rs_u \log\frac{1}{C\alpha}}. \tag{105}$$

Since λ is bounded away from 1, we may choose sufficiently small c_1 and α, so that the right hand side of (105) can be lower bounded by 0.9. This completes the first step.

Step 2. The part $\frac{s_u \log(ep/s_u)}{n\lambda^2}$ can be obtained from the rank-one argument spelled out in [12]. To be rigorous, consider the subset $T(\widetilde{U}^*)$ with $\widetilde{U}^* = \begin{bmatrix} I_{r-1} \\ 0 \end{bmatrix} \in O(2s_u/3, r-1)$. Restricting on the set $T(\widetilde{U}^*)$, the loss function is

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

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 [12, Appendix G] and choosing the appropriate constant, we have

$$\inf_{\widehat{U}} \sup_{T(\widetilde{U}^*)} \mathbb{P}\left(\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \geq C\frac{s_u \log(ep/s_u)}{n\lambda^2} \wedge c_0\right) \geq 0.9, \tag{106}$$

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 $\sup_{T(u_r^*) \cup T(\widetilde{U}^*)}$ on both sides of the inequality, and letting $\epsilon_1^2 = C_1\frac{rs_u}{n\lambda^2}$ in (105) and $\epsilon_2^2 = C_2\frac{s_u \log(ep/s_u)}{n\lambda^2} \wedge c_0$ in (106), we have

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

51

where we have used the identity $\sup_{\widetilde{U}\in T(u_r^*), u_r \in T(\widetilde{U}^*)} \left(f(u_r) + g(\widetilde{U}) \right) = \sup_{u_r \in T(\widetilde{U}^*)} f(u_r) + \sup_{\widetilde{U}\in T(u_r^*)} g(\widetilde{U})$. Careful readers may notice that we have assume sufficiently large r in Step 1. For r which is not sufficiently large, a similar rank-one argument as in Step 2 gives the desired lower bound. Thus, the proof is complete. $\qquad\square$

9.4   Proofs of technical lemmas

This section gathers the proofs of all technical results used in the above sections. The proofs are organized according to the order of their first appearance. To simplify notation, we denote $\widehat{\Sigma}_x^{(j)}$, $\widehat{\Sigma}_y^{(j)}$ and $\widehat{\Sigma}_{xy}^{(j)}$ by $\widehat{\Sigma}_x$, $\widehat{\Sigma}_y$ and $\widehat{\Sigma}_{xy}$ for $j \in \{0,1,2\}$ whenever there is no confusion from the context.

Let us first prove the claim (17) in Section 4.

Proof of (17). Remember that $\mathcal{O}_r = \{AB' : A \in O(p,r), B \in O(m,r)\}$ and $\mathcal{C}_r = \{G \in \mathbb{R}^{p\times m} : \|G\|_* \le r, \|G\|_{\mathrm{op}} \le 1\}$. Since $\mathcal{O}_r \subset \mathcal{C}_r$ and $\mathcal{C}_r$ is convex, we have $\mathrm{conv}(\mathcal{O}_r) \subset \mathcal{C}_r$. It is sufficient to show the other direction. For some $AB' \in \mathcal{O}_r$, we must have $-AB' \in \mathcal{O}_r$. Thus $0 = \frac{1}{2}(AB' - AB') \in \mathrm{conv}(\mathcal{O}_r)$. Hence, $\mathrm{conv}(\mathcal{O}_r) = \mathrm{conv}(\mathcal{O}_r \cup \{0\})$, and it is sufficient to prove $\mathcal{C}_r \subset \mathrm{conv}(\mathcal{O}_r \cup \{0\})$. For any $G \in \mathcal{C}_r$, it has SVD

$$G = \sum_{l=1}^{q} \lambda_l u_l v_l'.$$

Define

$$H_k = \sum_{l=k-1}^{k-1+r} u_l v_l', \text{ for } k = 2, ..., q, \quad \text{and} \quad H_1 = 0.$$

It is easy to see that $H_k \in \mathcal{O}_r \cup \{0\}$, for $k = 1, ..., q$, and

$$G = (1 - \lambda_1)H_0 + \sum_{k=2}^{q}(\lambda_{k-1} - \lambda_k)H_k.$$

Since $(1 - \lambda_1) + \sum_{k=2}^{q}(\lambda_{k-1} - \lambda_k) = 1$, we have $G \in \mathrm{conv}(\mathcal{O}_r \cup \{0\})$, and therefore $\mathcal{C}_r \subset \mathrm{conv}(\mathcal{O}_r \cup \{0\})$. Thus, the proof is complete. $\qquad\square$

Then we prove the lemmas in Section 6.

In order to prove Lemma 6.1, we need an auxiliary result.

Lemma 9.9. *Assume $\frac{1}{n}(s_u + s_v + \log(ep/s_u) + \log(em/s_v)) \le C_1$ for some constant $c > 0$. Then, for any $C' > 0$, there exists $C > 0$ only depending on C' such that*

$$\|U'\widehat{\Sigma}_x U - I\|_{\mathrm{op}} \vee \|(U'\widehat{\Sigma}_x U)^{1/2} - I\|_{\mathrm{op}} \le C\sqrt{\frac{1}{n}\left(s_u + \log\frac{ep}{s_u}\right)},$$

$$\|V'\widehat{\Sigma}_y V - I\|_{\mathrm{op}} \vee \|(V'\widehat{\Sigma}_y V)^{1/2} - I\|_{\mathrm{op}} \le C\sqrt{\frac{1}{n}\left(s_v + \log\frac{em}{s_v}\right)},$$

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

$$\|U'\widehat{\Sigma}_x U - I_r\|_{\mathrm{op}} = \|U'(\widehat{\Sigma}_x - \Sigma_x)U\|_{\mathrm{op}} = \|(U_{S_u*})'(\widehat{\Sigma}_{xS_uS_u} - \widehat{\Sigma}_{xS_uS_u})U_{S_u*}\|_{\mathrm{op}}$$

$$= \sup_{\|v\|=1} (U_{S_u*}v)'(\widehat{\Sigma}_{xS_uS_u} - \widehat{\Sigma}_{xS_uS_u})(U_{S_u*}v) \leq \|\Sigma_{xS_uS_u}^{1/2}U_{S_u*}\|_{\mathrm{op}}^2 \|\Sigma_{xS_uS_u}^{-1/2}\widehat{\Sigma}_{xS_uS_u}\Sigma_{xS_uS_u}^{-1/2} - I\|_{\mathrm{op}},$$

where $\|\Sigma_{xS_uS_u}^{1/2}U_{S_u*}\|_{\mathrm{op}}^2 \leq 1$ and $\|\Sigma_{xS_uS_u}^{-1/2}\widehat{\Sigma}_{xS_uS_u}\Sigma_{xS_uS_u}^{-1/2} - I\|_{\mathrm{op}}$ is bounded by the desired rate with high probability according to Lemma 16 in [16]. Lemma 15 in [16] 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. $\qquad\square$

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

$$\begin{aligned}
\|\Sigma_x^{1/2}(U - \widetilde{U})\|_{\mathrm{op}} &\leq \|\Sigma_x^{1/2}U\|_{\mathrm{op}}\|(U'\widehat{\Sigma}_x U)^{1/2} - I\|_{\mathrm{op}}\|(U'\widehat{\Sigma}_x U)^{-1/2}\|_{\mathrm{op}}, \\
\|\Sigma_y^{1/2}(V - \widetilde{V})\|_{\mathrm{op}} &\leq \|\Sigma_y^{1/2}V\|_{\mathrm{op}}\|(V'\widehat{\Sigma}_y V)^{1/2} - I\|_{\mathrm{op}}\|(V'\widehat{\Sigma}_y V)^{-1/2}\|_{\mathrm{op}}, \\
\|\widetilde{\Lambda} - \Lambda\|_{\mathrm{op}} &\leq \|(U'\widehat{\Sigma}_x U)^{1/2} - I\|_{\mathrm{op}}\|\Lambda(V'\widehat{\Sigma}_y V)^{1/2}\|_{\mathrm{op}} \\
&\quad + \|\Lambda\|_{\mathrm{op}}\|(V'\widehat{\Sigma}_y V)^{1/2} - I\|_{\mathrm{op}}.
\end{aligned}$$

Applying Lemma 9.9, the proof is complete. $\qquad\square$

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,

$$\|\widehat{\Sigma}_x^{1/2}\widetilde{A}\widehat{\Sigma}_y^{1/2}\|_{\mathrm{op}} \leq \|\widehat{\Sigma}_x^{1/2}\widetilde{U}\|_{\mathrm{op}}\|\widehat{\Sigma}_y^{1/2}\widetilde{V}\|_{\mathrm{op}} \leq 1. \tag{107}$$

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 $Q'Q = \widehat{\Sigma}_y^{1/2}V(V'\widehat{\Sigma}_y V)^{-1}V'\widehat{\Sigma}_y^{1/2}$, and

$$\mathsf{Tr}(Q'Q) = \mathsf{Tr}((V'\widehat{\Sigma}_y V)^{-1}(V'\widehat{\Sigma}_y V)) = r. \tag{108}$$

Combining (107) and (108), it is easy to see that all eigenvalues of $Q'Q$ are 1. Thus, we have $\|Q\|_* = r$ and $\|Q\|_{\mathrm{op}} = 1$. The proof is complete. $\qquad\square$

Proof of Lemma 6.3. Denote $F = [f_1, ..., f_r]$, $G = [g_1, ..., g_r]$ and $c_j = f_j'Eb_j$. By $\|E\|_{\mathrm{op}} \leq 1$, we have $|c_j| \leq 1$. The left hand side of (40) is lower bounded by

$$\langle FKG', FG' - E \rangle \geq \langle FDG', FG' - E \rangle - \|K - D\|_{\mathrm{F}}\|FG - E\|_{\mathrm{F}},$$

where

$$\langle FDG', FG' - E \rangle = \langle D, I - F'EG \rangle = \sum_{l=1}^{r} d_l(1 - c_l) \geq d_r \sum_{l=1}^{r}(1 - c_l).$$

The right hand side of (40) is

$$\frac{d_r}{2}\|FG' - E\|_{\mathrm{F}}^2 = \frac{d_r}{2}\Big(\|FG'\|_{\mathrm{F}}^2 + \|E\|_{\mathrm{F}}^2 - 2\,\mathsf{Tr}(F'EG)\Big)$$

$$\leq \frac{d_r}{2}\Big(\mathsf{Tr}(F'FG'G) + \|E\|_{\mathrm{op}}\|E\|_* - 2\sum_{j=1}^{r} c_j\Big)$$

$$\leq d_r \sum_{j=1}^{r}(1 - c_j).$$

This completes the proof. $\qquad\square$

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

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

The first term can be bounded by the desired rate by union bound and Bernstein's inequality [30, 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. $\qquad\square$

Proof of Lemma 6.6. By the property of least squares, we have

$$\inf_{W}\|F - GW\|_{\mathrm{F}}^2 \;=\; \|F - G(G'G)^- G'F\|_{\mathrm{F}}^2$$
$$=\; \|F - P_G F\|_{\mathrm{F}}^2$$
$$=\; r - \mathsf{Tr}(P_F P_G).$$

Since $\|P_F - P_G\|_{\mathrm{F}}^2 = 2r - 2\,\mathsf{Tr}(P_F P_G)$, the proof is complete. $\qquad\square$

Proof of Lemma 6.7. By the definition of U^*, we have $\Sigma_{xy}\widehat{V}^{(0)} = \Sigma_x U^*$. Thus,

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

Let us first bound $\max_{1\leq j\leq p}||[(\widehat{\Sigma}_x - \Sigma_x)U^*]_{j\cdot}||$. 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. Gaussian vectors distributed as $N(0, I_p)$. 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

$$\mathbb{P}\left\{\max_{1\leq j\leq p}||[(\widehat{\Sigma}_x - \Sigma_x)U^*]_{j\cdot}|| > t\right\}$$

$$\leq \sum_{j=1}^{p}\mathbb{P}\left\{\|\frac{1}{n}\sum_{i=1}^{n}(W_i^{(j)}(W_i^{(j)})' - \mathbb{E}W_i^{(j)}(W_i^{(j)})')\|_{\mathrm{op}} > t\right\}$$

$$\leq \sum_{j=1}^{p}\exp\left(C_1 r - C_2 n\min\left\{\frac{t}{\|\mathcal{W}^{(j)}\|_{\mathrm{op}}}, \frac{t^2}{\|\mathcal{W}^{(j)}\|_{\mathrm{op}}^2}\right\}\right), \tag{109}$$

where $\mathcal{W}^{(j)} = \mathbb{E}W_i^{(j)}(W_i^{(j)})'$, and we have used concentration inequality for sample covariance [30, Thm. 5.39]. Since $\|\mathcal{W}^{(j)}\|_{\mathrm{op}} \leq C_3$ for some constant C_3 only depending on M, (109) 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 $n^{-1}(r+\log p) \leq C_1$, $\max_{1\leq j\leq p}||[(\widehat{\Sigma}_x-\Sigma_x)U^*]_{j\cdot}|| \leq C\sqrt{\frac{r+\log p}{n}}$ with probability at least $1-\exp(-C'(r+\log p))$. Similar arguments lead to the bound of $\max_{1\leq j\leq p}||[(\widehat{\Sigma}_{xy} - \Sigma_{xy})\widehat{V}^{(0)}]_{j\cdot}||$. Let us sketch the proof. Note that we may write

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

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

and we have

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

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. $\qquad\square$

Finally, we prove the technical results in Section 9.

Proof of Lemma 9.1. Let us use the notation $\Delta = \widehat{U}\widehat{V}' - \widetilde{U}\widetilde{V}'$, $T_u = \widehat{S}_u \cup S_u$ and $T_v = \widehat{S}_v \cup S_v$, where $\widehat{S}_u = \mathrm{supp}(\widehat{U})$ and $\widehat{S}_v = \mathrm{supp}(\widehat{V})$. By the sparsity of Δ, we have

$$\|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_{\mathrm{F}}^2 = \mathsf{Tr}(\Sigma_x\Delta\Sigma_y\Delta') = \mathsf{Tr}(\Sigma_{xT_uT_u}\Delta_{T_uT_v}\Sigma_{yT_vT_v}(\Delta_{T_uT_v})') = \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_{\mathrm{F}}^2\,\mathsf{Tr}(KK'),$$

where

$$K = \|\Sigma_{xT_uT_u}^{1/2}\Delta_{T_uT_v}\Sigma_{yT_vT_v}^{1/2}\|_{\mathrm{F}}^{-1}\Sigma_{xT_uT_u}^{1/2}\Delta_{T_uT_v}\Sigma_{yT_vT_v}^{1/2} = \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_{\mathrm{F}}^{-1}\Sigma_{xT_uT_u}^{1/2}\Delta_{T_uT_v}\Sigma_{yT_vT_v}^{1/2},$$

so that $\|K\|_{\mathrm{F}} = 1$. Similarly, we have

$$\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 = \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_{\mathrm{F}}^2\,\mathsf{Tr}(\widehat{I}_x K \widehat{I}_y K'),$$

where

$$\widehat{I}_x = \Sigma_{xT_uT_u}^{-1/2}\widehat{\Sigma}_{xT_uT_u}\Sigma_{xT_uT_u}^{-1/2} \quad \text{and} \quad \widehat{I}_y = \Sigma_{yT_vT_v}^{-1/2}\widehat{\Sigma}_{yT_vT_v}\Sigma_{yT_vT_v}^{-1/2}.$$

Therefore,

$$\left| \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_{\mathrm{F}}^2 - \|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 \right| \le \|\Sigma_x^{1/2}\Delta\Sigma_y^{1/2}\|_{\mathrm{F}}^2 \left| \mathsf{Tr}(\widehat{I}_x K \widehat{I}_y K') - \mathsf{Tr}(KK') \right|. \qquad (110)$$

Note that

$$\begin{aligned}
&\left| \mathsf{Tr}(\widehat{I}_x K \widehat{I}_y K') - \mathsf{Tr}(KK') \right| \\
\le\;& \left| \mathsf{Tr}((\widehat{I}_x - I_{|T_u|})K\widehat{I}_y K') \right| + \left| \mathsf{Tr}(I_{|T_u|}K(\widehat{I}_y - I_{|T_v|})K') \right| \\
\le\;& \|(\widehat{I}_x - I_{|T_u|})K\|_{\mathrm{F}}\|\widehat{I}_y K'\|_{\mathrm{F}} + \|I_{|T_u|}K\|_{\mathrm{F}}\|(\widehat{I}_y - I_{|T_v|})K'\|_{\mathrm{F}} \\
\le\;& \|(\widehat{I}_x - I_{|T_u|})\|_{\mathrm{op}}\|\widehat{I}_y\|_{\mathrm{op}} + \|I_{|T_u|}\|_{\mathrm{op}}\|(\widehat{I}_y - I_{|T_v|})\|_{\mathrm{op}} \\
\le\;& C\left[\sqrt{\frac{s_u \log(ep/s_u)}{n}} + \sqrt{\frac{s_v \log(em/s_v)}{n}} \right],
\end{aligned}$$

with high probability, where we have used the fact that $\|K\|_{\mathrm{F}} = 1$ and the bounds of Lemma 12 in [16]. In view of (110), we have completed the proof. $\qquad\square$

Proof of Lemma 9.4. Let $T_v = \widehat{S}_v \cup S_v$, where $\widehat{S}_v = \mathrm{supp}(\widehat{V}^{(0)})$. First, let us bound $\|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}}$. Since $(\widehat{V}^{(0)})'\widehat{\Sigma}_y\widehat{V}^{(0)} = I_r$, we have

$$
\begin{aligned}
\|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}} &\leq \|\Sigma^{1/2}_{yT_vT_v}\widehat{\Sigma}^{-1/2}_{yT_vT_v}\|_{\mathrm{op}}\|\widehat{\Sigma}^{1/2}_{yT_vT_v}\widehat{V}^{(0)}\|_{\mathrm{op}} \leq \|\Sigma^{1/2}_{yT_vT_v}\widehat{\Sigma}^{-1/2}_{yT_vT_v}\|_{\mathrm{op}} \\
&= \|\Sigma^{1/2}_{yT_vT_v}\widehat{\Sigma}^{-1}_{yT_vT_v}\Sigma^{1/2}_{yT_vT_v}\|^{1/2}_{\mathrm{op}} = \sigma_{\min}(\Sigma^{-1/2}_{yT_vT_v}\widehat{\Sigma}_{yT_vT_v}\Sigma^{-1/2}_{yT_vT_v})^{-1/2} \\
&\leq \left(1 - \|\Sigma^{-1/2}_{yT_vT_v}\widehat{\Sigma}_{yT_vT_v}\Sigma^{-1/2}_{yT_vT_v} - I\|_{\mathrm{op}}\right)^{-1/2} \\
&\leq 1 + C\sqrt{\frac{s_u\log(ep/s_u)}{n}},
\end{aligned}
$$

with probability at least $1 - \exp(-C's_u\log(ep/s_u))$, where the last inequality is by Lemma 12 of [16]. Hence,

$$
\begin{aligned}
\|(\widehat{V}^{(0)})'\Sigma_y\widehat{V}^{(0)} - I\|_{\mathrm{op}} &= \|(\widehat{V}^{(0)})'(\Sigma_y - \widehat{\Sigma}_y)\widehat{V}^{(0)}\|_{\mathrm{op}} \\
&= \|(\widehat{V}^{(0)}_{T_v*})'(\Sigma_{yT_vT_v} - \widehat{\Sigma}_{yT_vT_v})\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}} \\
&\leq \|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|^2_{\mathrm{op}}\|\Sigma^{-1/2}_{yT_vT_v}\widehat{\Sigma}_{yT_vT_v}\Sigma^{-1/2}_{yT_vT_v} - I\|_{\mathrm{op}} \\
&\leq 4C\sqrt{\frac{s_u\log(ep/s_u)}{n}},
\end{aligned}
$$

with probability at least $1 - \exp(-C's_u\log(ep/s_u))$. The proof is completed by realizing $\|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}} = \|\Sigma^{1/2}_y\widehat{V}^{(0)}\|_{\mathrm{op}}$. $\qquad\square$

Proof of Lemma 9.5. Let $T_u = \widehat{S}_u \cup S_u$, where $\widehat{S}_u = \mathrm{supp}(\widehat{U})$. Using the definition of Frobenius norm, we have

$$
\begin{aligned}
\left|\|\Sigma^{1/2}_x\Delta\|^2_{\mathrm{F}} - \|\widehat{\Sigma}^{1/2}_x\Delta\|^2_{\mathrm{F}}\right| &= \left|\mathsf{Tr}(\Delta'(\widehat{\Sigma}_x - \Sigma_x)\Delta)\right| \\
&= \left|\mathsf{Tr}((\Delta_{T_u*})'(\widehat{\Sigma}_{xT_uT_u} - \Sigma_{xT_uT_u})\Delta_{T_u*})\right| \\
&\leq \|\Sigma_{xT_uT_u}\Delta_{T_u*}\|^2_{\mathrm{F}}\|\Sigma^{-1/2}_{xT_uT_u}\widehat{\Sigma}_{xT_uT_u}\Sigma^{-1/2}_{xT_uT_u} - I\|_{\mathrm{op}} \\
&\leq C\sqrt{\frac{s_u\log(ep/s_u)}{n}}\|\Sigma^{1/2}_x\Delta\|^2_{\mathrm{F}},
\end{aligned}
$$

with high probability, where we have used $\|\Sigma_{xT_uT_u}\Delta_{T_u*}\|^2_{\mathrm{F}} = \|\Sigma^{1/2}_x\Delta\|^2_{\mathrm{F}}$ and Lemma 12 in [16] in the last inequality. After rearrangement, the proof is complete. $\qquad\square$

Proof of Lemma 9.6. In this proof, $\widehat{\Sigma}_x$ is constructed from $\mathcal{D}_0$ and $\widehat{\Sigma}_y$ is constructed from $\mathcal{D}_1$. We use the notation $T_u = S_u \cup \widehat{S}_u$ and $T_v = S_v \cup \widehat{S}_v$, where $\widehat{S}_u = \mathrm{supp}(\widehat{U})$ and $\widehat{S}_v = \mathrm{supp}(\widehat{V}^{(0)})$. Note that T_u depends on $\mathcal{D}_1$ and T_v depends on $\mathcal{D}_0$. We first condition on $\mathcal{D}_0$, and then we

have

$$
\left| \mathsf{Tr}\left(\Delta'(\widehat{\Sigma}_{xy} - \Sigma_{xy})\widehat{V}^{(0)} \right) \right|
$$

$$
= \left| \langle \widehat{\Sigma}_{xyT_uT_v} - \Sigma_{xyT_uT_v}, \Delta'_{T_u*}(\widehat{V}^{(0)}_{T_v*})' \rangle \right|
$$

$$
\leq \|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}}\|\Sigma^{1/2}_{xT_uT_u}\Delta_{T_u*}\|_{\mathrm{F}} \left| \langle \Sigma^{-1/2}_{xT_uT_u}(\widehat{\Sigma}_{xyT_uT_v} - \Sigma_{xyT_uT_v})\Sigma^{-1/2}_{yT_vT_v}, K_{T_u} \rangle \right|
$$

$$
\leq \|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}}\|\Sigma^{1/2}_{xT_uT_u}\Delta_{T_u*}\|_{\mathrm{F}} \sup_T \left| \langle \Sigma^{-1/2}_{xTT}(\widehat{\Sigma}_{xyTT_v} - \Sigma_{xyTT_v})\Sigma^{-1/2}_{yT_vT_v}, K_T \rangle \right|
$$

where T ranges over all subsets with cardinality bounded by $2s_u$, and for each such T, $K_T = \|\Sigma^{1/2}_{xTT}\Delta'_{T*}(\widehat{V}^{(0)}_{T_v*})'\Sigma^{1/2}_{yT_vT_v}\|_{\mathrm{F}}^{-1}\Sigma^{1/2}_{xTT}\Delta'_{T*}(\widehat{V}^{(0)}_{T_v*})'\Sigma^{1/2}_{yT_vT_v}$ satisfying $\|K_T\|_{\mathrm{F}} = 1$. We do not put T_v in the subscript of K because conditioning on $\mathcal{D}_0$, T_v is fixed. For each T, we can use Lemma 7 in [16] to bound $\left| \langle \Sigma^{-1/2}_{xTT}(\widehat{\Sigma}_{xyTT_v} - \Sigma_{xyTT_v})\Sigma^{-1/2}_{yT_vT_v}, K_T \rangle \right|$. A direct union bound argument leads to

$$
\sup_T \left| \langle \Sigma^{-1/2}_{xTT}(\widehat{\Sigma}_{xyTT_v} - \Sigma_{xyTT_v})\Sigma^{-1/2}_{yT_vT_v}, K_T \rangle \right| \leq C\sqrt{\frac{rs_u + s_u\log(ep/s_u)}{n}},
$$

with probability at least $1-\exp\left(-C'(s_u\log(ep/s_u) + rs_u)\right)$. By Lemma 9.4, we have $\|\Sigma^{1/2}_{yT_vT_v}\widehat{V}^{(0)}_{T_v*}\|_{\mathrm{op}} = \|\Sigma^{1/2}_y\widehat{V}^{(0)}\|_{\mathrm{op}} \leq 2$ with high probability. Finally, observing that $\|\Sigma^{1/2}_{xT_uT_u}\Delta_{T_u*}\|_{\mathrm{F}} = \|\Sigma^{1/2}_x\Delta\|_{\mathrm{F}}$, we have completed the proof. $\qquad\square$

Proof of Lemma 9.7. The proof is very similar to that of Lemma 9.6, and is thus omitted. $\quad\square$

Proof of Proposition 9.2. Let the singular value decomposition of U be $U = \Theta RH'$. Then we have $HR\Theta'\Sigma_x\Theta RH' = U'\Sigma_x U = I$, from which we derive $\Theta'\Sigma_x\Theta = R^{-2}$. Using Lemma 6.6, we have

$$
\begin{aligned}
\|P_{\widehat{U}} - P_U\|_{\mathrm{F}} &= \sqrt{2}\inf_W \|\widehat{U}W - \Theta\|_{\mathrm{F}} \\
&\leq \sqrt{2}\inf_W \|\widehat{U}WHR^{-1} - \Theta RH'HR^{-1}\|_{\mathrm{F}} \\
&\leq \sqrt{2}\inf_W \|\widehat{U}W - U\|_{\mathrm{F}}\|R^{-1}\|_{\mathrm{op}} \\
&\leq \sqrt{2}M^{1/2}\inf_W \|\Sigma^{1/2}_x(\widehat{U}W - U)\|_{\mathrm{F}}\|\Theta'\Sigma_x\Theta\|^{1/2}_{\mathrm{op}} \\
&\leq \sqrt{2}M\inf_W \|\Sigma^{1/2}_x(\widehat{U}W - U)\|_{\mathrm{F}} \\
&\leq \sqrt{2}M\inf_{W\in O(r,r)} \|\Sigma^{1/2}_x(\widehat{U}W - U)\|_{\mathrm{F}}.
\end{aligned}
$$

Finally, by $\|\Sigma^{1/2}_x(\widehat{U}W - U)\|^2_{\mathrm{F}} = \mathsf{Tr}((\widehat{U}W - U)'\Sigma_x(\widehat{U}W - U))$, the proof is complete. $\qquad\square$

10 Implementation of (18)

To implement the convex programming (18), we turn to the Alternating Direction Method of Multipliers (ADMM) [14, 10]. In the rest of this section, we write $\widehat{\Sigma}_x$ and $\widehat{\Sigma}_y$ for $\widehat{\Sigma}^{(0)}_x$ and $\widehat{\Sigma}^{(0)}_y$ for notational convenience.

First, note that (18) can be rewritten as

$$
\begin{aligned}
\text{minimize} \quad & f(F) + g(G), \\
\text{subject to} \quad & \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G = 0,
\end{aligned}
\tag{111}
$$

where

$$
f(F) = -\langle \widehat{\Sigma}_{xy}, F \rangle + \rho \|F\|_1, \tag{112}
$$

$$
g(G) = \infty \mathbf{1}_{\{\|G\|_* > r\}} + \infty \mathbf{1}_{\{\|G\|_{\mathrm{op}} > 1\}}. \tag{113}
$$

Thus, the augmented Lagrangian form of the problem is

$$
\mathcal{L}_\eta(F, G, H) = f(F) + g(G) + \langle H, \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G \rangle + \frac{\eta}{2} \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G\|_{\mathrm{F}}^2. \tag{114}
$$

Following the generic algorithm spelled out in Section 3 of [10], 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:

$$
F^{k+1} = \underset{F}{\operatorname{argmin}} \, \mathcal{L}_\eta(F, G^k, H^k), \tag{115}
$$

$$
G^{k+1} = \underset{G}{\operatorname{argmin}} \, \mathcal{L}_\eta(F^{k+1}, G, H^k), \tag{116}
$$

$$
H^{k+1} = H^k + \eta(\widehat{\Sigma}_x^{1/2} F^{k+1} \widehat{\Sigma}_y^{1/2} - G^{k+1}). \tag{117}
$$

The algorithm iterates over (115) – (117) till some convergence criterion is met. It is clear that the update (117) for the dual variable H is easy to calculate. Moreover the updates (115) and (116) can be solved easily and have explicit meaning in giving solution to sparse CCA. We are going to show that (115) can be viewed as a Lasso problem. Thus, this step targets at the sparsity of the matrix UV'. The update (116) 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 the updates for F and G.

First, we note that (115) is equivalent to

$$
\begin{aligned}
F^{k+1} &= \underset{F}{\operatorname{argmin}} \, f(F) + \langle H^k, \widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} \rangle + \frac{\eta}{2} \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - G^k\|_{\mathrm{F}}^2 \\
&= \underset{F}{\operatorname{argmin}} \, \frac{\eta}{2} \|\widehat{\Sigma}_x^{1/2} F \widehat{\Sigma}_y^{1/2} - (G^k - \frac{1}{\eta} H^k + \frac{1}{\eta} \widehat{\Sigma}_x^{-1/2} \widehat{\Sigma}_{xy} \widehat{\Sigma}_y^{-1/2})\|_{\mathrm{F}}^2 + \rho \|F\|_1. \tag{118}
\end{aligned}
$$

Thus, it is clear that the update of F in (115) can be viewed as a Lasso problem as summarized in the following proposition. Here and after, for any positive semi-definite matrix A, $A^{-1/2}$ denotes the principal square root of its pseudo-inverse.

Proposition 10.1. *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 following standard Lasso problem*

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

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

Remark 10.1. It is worth mentioning that the vectorized formulation in Proposition 10.1 is for illustration only. In practice, we solve the problem in (118) directly, since the vectorized version, especially the Kronecker product, would great increase the computation cost. The solver to (118) can be easily implemented in standard software packages for convex programming, such as TFOCS [5].

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 (116) is equivalent to

$$
\begin{aligned}
G^{k+1} &= \operatorname*{argmin}_{G} g(G) - \langle H^k, G \rangle + \frac{\eta}{2} \|\widehat{\Sigma}_x^{1/2} F^{k+1} \widehat{\Sigma}_y^{1/2} - G\|_{\mathrm{F}}^2 \\
&= \operatorname*{argmin}_{G} \frac{\eta}{2} \|G - (\frac{1}{\eta} H^k + \widehat{\Sigma}_x^{1/2} F^{k+1} \widehat{\Sigma}_y^{1/2})\|_{\mathrm{F}}^2 \\
&\qquad\qquad + \infty \mathbf{1}_{\{\|G\|_* > r\}} + \infty \mathbf{1}_{\{\|G\|_{\mathrm{op}} > 1\}} \\
&= \operatorname*{argmin}_{G} \|G - (\frac{1}{\eta} H^k + \widehat{\Sigma}_x^{1/2} F^{k+1} \widehat{\Sigma}_y^{1/2})\|_{\mathrm{F}}^2 \\
&\qquad\qquad + \infty \mathbf{1}_{\{\|G\|_* > r\}} + \infty \mathbf{1}_{\{\|G\|_{\mathrm{op}} > 1\}}.
\end{aligned}
\tag{119}
$$

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

Proposition 10.2. *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 [31]. 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 [31]. The asymmetry of the current problem does not matter since it is orthogonally invariant. $\square$

Here and after, we call the operation in Proposition 10.2 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 (18) is implemented as Algorithm 1.

11   Numerical Studies

This section presents numerical results demonstrating the competitive finite sample performance of the proposed adaptive estimation procedure CoLaR on simulated datasets.

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 $\widehat{A}$.

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

   repeat

2   Update F^{k+1} as in (115)   (Lasso) ;

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

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

5       $k \leftarrow k + 1$;

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

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

Simulation settings We consider three simulation settings. In all these settings, we set $p = m$, $\Sigma_x = \Sigma_y = \Sigma$, and $r = 2$ with $\lambda_1 = 0.9$ and $\lambda_2 = 0.8$. Moreover, the nonzero rows of both U and V are set at $\{1, 6, 11, 16, 21\}$. The values at the nonzero coordinates are obtained from normalizing (with respect to Σ) random numbers drawn from the uniform distribution on the finite set $\{-2, 1, 0, 1, 2\}$. The choices of Σ in the three settings are as follows:

1. `Identity`: $\Sigma = I_p$.

2. `Toeplitz`: $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = 0.3^{|i-j|}$ for all $i, j \in [p]$. In other words, Σ_x and Σ_y are Toeplitz matrices.

3. `SparseInv`: $\Sigma = (\sigma_{ij}^0 / \sqrt{\sigma_{ii}^0 \sigma_{jj}^0})$. We set $\Sigma^0 = (\sigma_{ij}^0) = \Omega^{-1}$ where $\Omega = (\omega_{ij})$ 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.

In all three settings, we normalize the variance of each coordinate to be one.

Implementation details The proposed CoLaR estimator in Section 4.1 has two stages. The convex program (18) in the first stage can be solved via an ADMM algorithm [10]. The details of the ADMM approach are presented in Section 10. The optimization problem (19) in the second stage can be solved by a standard group-Lasso algorithm [40].

    In all numerical results reported in this section, we used the same penalty level $\rho = 0.55\sqrt{\log(p+m)/n}$ in (18) and we used $\eta = 2$ in (117). In (19), we used five-fold cross validation to select a common penalty parameter $\rho_u = \rho_v = b\sqrt{(r + \log p)/n}$. In particular,

for $l = 1, \ldots, 5$, we use one fold of the data as the test set $(X_{(l)}^{\text{test}}, Y_{(l)}^{\text{test}})$ and the other four folds as the training set $(X_{(l)}^{\text{train}}, Y_{(l)}^{\text{train}})$. For any choice of b, we solved (19) on $(X_{(l)}^{\text{train}}, Y_{(l)}^{\text{train}})$ to obtain estimates $(\widehat{U}_{(l)}, \widehat{V}_{(l)})$. Then we computed the sum of canonical correlations between $X_{(l)}^{\text{test}}\widehat{U}_{(l)} \in \mathbb{R}^{n \times r}$ and $Y_{(l)}^{\text{test}}\widehat{V}_{(l)} \in \mathbb{R}^{n \times r}$ to obtain $\text{CV}_l(b)$. Finally, $\text{CV}(b) = \sum_{l=1}^{5} \text{CV}_l(b)$. Among all the candidate penalty parameters, we select the b value such that $\text{CV}(b)$ is maximized. The candidate b values used in the simulation below are $\{0.5, 1, 1.5, 2\}$. Throughout the simulation, we used all the sample $\{(X_i, Y_i)\}_{i=1}^{n}$ to form the sample covariance matrices used in (18) – (20).

In addition to the performance of CoLaR, we also report that of the method proposed in [38] (denoted by PMA here and on). The PMA seeks the solution to the optimization problem

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

The solution is used to estimate the first canonical pair $(\widehat{u}_1, \widehat{v}_1)$. Then 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 gives the estimator of 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 (function `CCA` in the `PMA` package in R) by the authors of [38]. To remove undesired amplification of error caused by normalization, we renormalized each individual $\widehat{u}_j$ with respect to $\widehat{\Sigma}_x$ and each individual $\widehat{v}_j$ with respect to $\widehat{\Sigma}_y$ before calculating the error under the loss (7). For each simulated dataset, we set the sparsity penalty parameters `penaltyx` and `penaltyz` of the function `CCA` at each of the eleven different values $\{0.6^l : l = 0, 1, \ldots, 10\}$ and only the smallest estimation error out of all eleven trials was used to compute the error reported in the tables below.

Results Tables 1 – 3 report, in each of the three settings, the medians of the prediction errors of CoLaR and PMA out of 100 repetitions for four different configurations of (p, m, n) values.

In each table, the columns U-PMA and V-PMA report the medians of the smallest estimation errors out of the eleven trials on each simulated dataset. The columns U-init and V-init report the median estimation errors of the renormalized r left singular vectors and right singular vectors of the solutions to the initialization step (18), where the renormalization is the same as in (20) and in both (18) and renormalization we used all the n pairs of observations. Last but not least, the columns U-CoLaR and V-CoLaR report the median estimation errors of the CoLaR estimators where both stages were carried out.

In all simulation settings, both the renormalized initial estimators and the CoLaR estimators consistently outperform PMA. Comparing the last four columns within each table, we also find that the CoLaR estimators with both stages carried out significantly improve over the renormalized initial estimators, which is in accordance with our theoretical results in Section 4.

In summary, the proposed method delivers consistent and competitive performance in all three covariance settings across all dimension and sample size configurations, and its behavior agrees well with the theory.

(p, m, n)	U-PMA	V-PMA	U-init	V-init	U-CoLaR	V-CoLaR
(300, 300, 200)	2.1316	2.1297	0.2653	0.1712	0.0498	0.0646
(600, 600, 200)	3.4154	3.3584	0.3167	0.2087	0.0671	0.0776
(300, 300, 500)	0.2683	0.2701	0.1207	0.0665	0.0135	0.0159
(600, 600, 500)	2.0335	2.0368	0.1448	0.0817	0.0166	0.0203

Table 1: Prediction errors (`Identity`): Median in 100 repetitions.

(p, m, n)	U-PMA	V-PMA	U-init	V-init	U-CoLaR	V-CoLaR
(300, 300, 200)	2.1853	2.1840	0.2885	0.1706	0.0511	0.0601
(600, 600, 200)	3.4247	3.4852	0.3236	0.2004	0.0638	0.0764
(300, 300, 500)	0.2358	0.2191	0.1202	0.0664	0.0135	0.0166
(600, 600, 500)	2.1214	2.0889	0.1408	0.0811	0.0176	0.0209

Table 2: Prediction errors (`Toeplitz`): Median in 100 repetitions.

(p, m, n)	U-PMA	V-PMA	U-init	V-init	U-CoLaR	V-CoLaR
(300, 300, 200)	2.9697	2.9619	0.5552	0.5718	0.1568	0.1194
(600, 600, 200)	4.6908	4.3339	0.5596	0.6133	0.2123	0.1572
(300, 300, 500)	2.3967	2.0620	0.2695	0.1917	0.0242	0.0219
(600, 600, 500)	2.8707	2.8609	0.3068	0.2368	0.0338	0.0271

Table 3: Prediction errors (`SparseInv`): Median in 100 repetitions.