The Annals of Statistics
2017, Vol. 45, No. 5, 2074–2101
DOI: 10.1214/16-AOS1519

SPARSE CCA: ADAPTIVE ESTIMATION AND COMPUTATIONAL BARRIERS

BY CHAO GAO[1], ZONGMING MA[2] AND HARRISON H. ZHOU[1]

University of Chicago, University of Pennsylvania and Yale University

Canonical correlation analysis is a classical technique for exploring the relationship between two sets of variables. It has important applications in analyzing high dimensional datasets originated from genomics, imaging and other fields. This paper considers adaptive minimax and computationally tractable estimation of leading sparse canonical coefficient vectors in high dimensions. Under a Gaussian canonical pair model, we first establish separate minimax estimation rates for canonical coefficient vectors of each set of random variables under no structural assumption on marginal covariance matrices. Second, we propose a computationally feasible estimator to attain the optimal rates adaptively under an additional sample size condition. Finally, we show that a sample size condition of this kind is needed for any randomized polynomial-time estimator to be consistent, assuming hardness of certain instances of the planted clique detection problem. As a byproduct, we obtain the first computational lower bounds for sparse PCA under the Gaussian single spiked covariance model.

1. Introduction. Canonical correlation analysis (CCA) [20] is a classical and important tool 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 coefficient vectors*) that solve

$$\begin{aligned} &\max_{a,b} \quad a'\Sigma_{xy}b, \\ &\text{subject to} \quad a'\Sigma_x a = b'\Sigma_y b = 1, \\ &\qquad\qquad a'\Sigma_x u_l = b'\Sigma_y v_l = 0, \qquad \forall 0 \le l \le j-1, \end{aligned} \tag{1}$$

where $\Sigma_x = \mathrm{Cov}(X)$, $\Sigma_y = \mathrm{Cov}(Y)$, $\Sigma_{xy} = \mathrm{Cov}(X, Y)$, $u_0 = 0$ 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 jth pair of *canonical variates*. This technique has been widely used in

Received August 2015; revised September 2016.

[1]Supported in part by NSF Grants DMS-12-09191 and DMS-15-07511. The research was mainly conducted when C. Gao was at Yale University.

[2]Supported in part by NSF Career Award DMS-1352060 and a Sloan Research Fellowship.

MSC2010 subject classifications. Primary 62H12; secondary 62C20.

Key words and phrases. Convex programming, group-Lasso, minimax rates, computational complexity, planted clique, sparse CCA (SCCA), sparse PCA (SPCA).

various scientific fields to explore the relationship 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 been well understood that classical CCA breaks down in this regime [4, 16, 21]. Motivated by genomics, neuroimaging and other applications, people have become interested in seeking sparse leading canonical coefficient vectors. Various estimation procedures imposing sparsity on canonical coefficient vectors have been developed in the literature, which are usually termed *sparse CCA*. See, for example, [3, 19, 23, 29, 33, 36, 37].

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 [13]. In particular, suppose there are r pairs of canonical coefficient vectors (and canonical variates) among the two sets of variables, then the model reparameterizes the cross-covariance matrix as

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

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

The goal of the present paper is to study three fundamental questions in sparse CCA: (1) What are the minimax rates for estimating the canonical coefficient vectors 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?

1.1. *Main contributions.* We now introduce the main contributions of the present paper from three different viewpoints as suggested by the three questions we have raised.

Separate minimax rates. The joint loss $\|\widehat{U}\widehat{V}' - UV'\|_{\mathrm{F}}^2$ studied by [16] characterizes the joint estimation error of both U and V. In this paper, we provide a finer analysis by studying individual estimation errors of U and V 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 minimax rate of convergence in estimating U depends only on 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 of the rates of estimating U and V.

Adaptive estimation. As pointed out in [13] and [16], sparse CCA is a more difficult problem than the well-studied sparse PCA. A naive application of sparse PCA algorithm to sparse CCA can lead to inconsistent results [13]. The additional difficulty in sparse CCA mainly comes from the presence of the nuisance parameters Σ_x and Σ_y, which cannot be estimated consistently in a high-dimensional regime in general. Therefore, our goal is to design an estimator that is adaptive to both the nuisance parameters and the sparsity levels. Under the canonical pair model, we propose a computationally efficient algorithm. The algorithm has two stages. In the first stage, we propose a convex program 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 program can be efficiently solved by the Alternating Direction Method with Multipliers (ADMM) [10, 14]. Based on the output of the first stage, we formulate a sparse linear regression problem in the second stage to improve estimation accuracy, and the final estimator $\widehat{U}$ and $\widehat{V}$ can be obtained via a group-Lasso algorithm [38]. Under the sample size condition that

$$n \geq Cs_u s_v \log(p+m)/\lambda_r^2 \tag{4}$$

for some sufficiently large constant $C > 0$, we show $\widehat{U}$ and $\widehat{V}$ recover the true canonical coefficient matrices U and V within optimal error rates adaptively with high probability. A Matlab implementation of the proposed estimator is available at http://www-stat.wharton.upenn.edu/~zongming/software/SCCALab.zip.

Computational lower bound. We require the sample size condition (4) for the adaptive procedure to achieve optimal rates of convergence. Assuming hardness of certain instances of the Planted Clique detection problem, we provide a computational lower bound to show that a condition of this kind is unavoidable for any computationally feasible estimation procedure to achieve consistency. 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 has been imposed in the sparse PCA literature [12, 22, 25, 32], 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. Berthet and Rigollet [6] showed that if there existed a polynomial-time algorithm for a generalized sparse PCA detection problem while such a condition is violated, 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 in [6] to include all multivariate distributions whose quadratic forms satisfy certain uniform tail probability bounds and so the distributions need not be Gaussian or having a spiked covariance structure [22]. The same remark also applies to the subsequent work on sparse PCA estimation [34]. 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 Gaussian single spiked covariance model.

1.2. *Organization.* After an introduction to notation, 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 separate minimax rates for estimating U and V. Section 4 proposes a two-stage adaptive estimator that is shown to be minimax rate optimal under an additional sample size condition. Section 5 shows a condition of this kind is necessary for all randomized polynomial-time estimator to achieve consistency by establishing new computational lower bounds for sparse PCA and sparse CCA. Section 6 presents proofs of theoretical results in Section 4. Implementation details of the adaptive procedure, numerical studies, additional proofs and further discussion are deferred to the supplement [17].

1.3. *Notation.* For any $t \in \mathbb{Z}_+$, $[t]$ denotes the set $\{1, 2, \ldots, t\}$. For any set S, $|S|$ denotes its cardinality and S^c its complement. For any event E, $\mathbf{1}_{\{E\}}$ denotes its indicator function. For any $a, b \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer no smaller than a, $\lfloor a \rfloor$ the largest integer no larger than a, $a \vee b = \max(a, b)$ 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 $A = (a_{ij}) \in \mathbb{R}^{p \times k}$, $A_{i\cdot}$ denotes its ith row and $\operatorname{supp}(A) = \{i \in [p] : \|A_{i\cdot}\| > 0\}$, the index set of nonzero rows, is called its support. For any subset $J \subset [p] \times [k]$, $A_J = (a_{ij}\mathbf{1}_{\{(i,j)\in J\}}) \in \mathbb{R}^{p \times k}$ is obtained by keeping all entries in J and replacing all entries in J^c with zeros. We write $A_{J_1 J_2}$ for $A_{J_1 \times J_2}$ and $A_{(J_1 J_2)^c}$ for $A_{(J_1 \times J_2)^c}$. Notice that $A_{J_1 *} = A_{J_1 \times [k]} \in \mathbb{R}^{p \times k}$ while $A_{J_1 \cdot}$ stands for the corresponding nonzero submatrix of size $|J_1| \times k$. In addition, $P_A \in \mathbb{R}^{p \times p}$ stands for the projection matrix onto the column space of A, $O(p, k)$ denotes the set of all $p \times k$ orthogonal matrices and $O(k) = O(k, k)$. Furthermore, $\sigma_i(A)$ stands for the ith largest singular value of A and $\sigma_{\max}(A) = \sigma_1(A)$,

$\sigma_{\min}(A) = \sigma_{p\wedge k}(A)$. The Frobenius norm and the operator norm of A are $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j} a_{ij}^2}$ and $\|A\|_{\mathrm{op}} = \sigma_1(A)$, respectively. The l_1 norm and the nuclear norm are $\|A\|_1 = \sum_{ij} |a_{ij}|$ and $\|A\|_* = \sum_i \sigma_i(A)$, respectively. If A is a square matrix, its trace is $\mathsf{Tr}(A) = \sum_i a_{ii}$. For two square matrices A and B, we write $A \preceq B$ if $B - A$ is positive semidefinite. 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 $\langle A, B\rangle = \mathsf{Tr}(A'B)$. Given a random element X, $\mathcal{L}(X)$ denotes its probability distribution. The symbol C and its variants C_1, C', etc. are generic constants and may vary from line to line, unless otherwise specified. The symbols $\mathbb{P}$ 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 coefficient vectors are sparse. 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 [12, 32] 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:

$$\begin{aligned} &1.\quad U \in \mathbb{R}^{p\times r} \text{ and } V \in \mathbb{R}^{m\times r} \text{ with } \big|\mathrm{supp}(U)\big| \le s_u \text{ and } \big|\mathrm{supp}(V)\big| \le s_v; \\ &2.\quad \sigma_{\min}(\Sigma_x) \wedge \sigma_{\min}(\Sigma_y) \ge M^{-1} \text{ and } \sigma_{\max}(\Sigma_x) \vee \sigma_{\max}(\Sigma_y) \le M; \\ &3.\quad \lambda_r \ge \lambda \text{ and } \lambda_1 \le 1 - M^{-1}. \end{aligned} \tag{5}$$

The probability space we consider is

$$\begin{aligned} &\mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M) \\ &\quad = \big\{\mathcal{L}\big((X_1', Y_1')', \ldots, (X_n', Y_n')'\big) : (X_i', Y_i')' \overset{\text{i.i.d.}}{\sim} N_{p+m}(0, \Sigma) \\ &\qquad \text{with } \Sigma \in \mathcal{F}(s_u, s_v, p, m, r, \lambda; M)\big\}, \end{aligned} \tag{6}$$

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.

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 coefficient vectors 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:

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

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. 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 coefficient vectors 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

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

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. An attractive feature of the loss function (7) is its invariance with respect to linear transformations on X and Y. If each X_i and Y_i are transformed to RX_i and TY_i for any invertible matrices R and T, then the canonical coefficient matrices become $R^{-1}U$ and $T^{-1}V$, respectively. Thus, the value of the loss (7) does not change if we replace $\widehat{U}$ and U with $R^{-1}\widehat{U}$ and $R^{-1}U$. Careful readers might realize that an arbitrary linear transform R may change the sparsity pattern of U. However, if R is diagonal, the sparsity pattern is preserved. So the sparse CCA problem paired with the loss (7) is insensitive to any scale change of the variables.

A related loss function is $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$ measuring the difference between two subspaces. By Proposition 9.2 in the supplementary material [17], we have $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 \le CL(\widehat{U}, U)$ for some constant $C > 0$ only depending on M. Moreover, the loss (7) contains more information on the difference between U and $\widehat{U}$. To see this, let $\Sigma_x = I_p$, $U \in O(p, r)$ and $\widehat{U} = 2U$. Then $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2 = 0$, while $L(\widehat{U}, U) = \inf_{W \in O(r)} \|\Sigma_x^{1/2}(\widehat{U}W - U)\|_{\mathrm{F}}^2 = \inf_{W \in O(r)}(5r - \mathsf{Tr}(W)) = r > 0$. In this paper, we will focus on the loss $L(\widehat{U}, U)$ while providing brief remarks on results for $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$.

3. Minimax rates. 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, that is, $n = 3n_0$ 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:

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

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

$$\begin{aligned} &\min_{L \in \mathbb{R}^{p\times r}} \operatorname{Tr}\bigl(L' \widehat{\Sigma}_x^{(1)} L\bigr) - 2 \operatorname{Tr}\bigl(L' \widehat{\Sigma}_{xy}^{(1)} \widehat{V}^{(0)}\bigr) \\ &\text{subject to} \quad \bigl|\operatorname{supp}(L)\bigr| \le s_u. \end{aligned} \tag{10}$$

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

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

The purpose of sample splitting employed in the above procedure is to facilitate the proof.

Theory and discussion. The program (9) was first proposed in [16] as a sparsity constrained version of the classical CCA formulation. However, the resulting estimator will have a convergence rate that involves the sparsity level s_v and the ambient dimension m of the V matrix [16], Theorem 1, which is sub-optimal. The second stage in the procedure is thus proposed to further pursue the optimal estimation rates. 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

$$\begin{aligned} U\Lambda &= \operatorname*{argmin}_{L \in \mathbb{R}^{p\times r}} \mathbb{E} \bigl\| Y'V - X'L \bigr\|^2 \\ &= \operatorname*{argmin}_{L \in \mathbb{R}^{p\times r}} \operatorname{Tr}\bigl(L' \Sigma_x L\bigr) - 2 \operatorname{Tr}\bigl(L' \Sigma_{xy} V\bigr) + \operatorname{Tr}\bigl(V' \Sigma_y V\bigr) \\ &= \operatorname*{argmin}_{L \in \mathbb{R}^{p\times r}} \operatorname{Tr}\bigl(L' \Sigma_x L\bigr) - 2 \operatorname{Tr}\bigl(L' \Sigma_{xy} V\bigr), \end{aligned} \tag{12}$$

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 Euclidean norm, and the last equality holds since $\mathrm{Tr}(V'\Sigma_y V)$ does not involve the argument to be optimized over. In fact, from the canonical pair model, one can easily derive a regression interpretation of CCA, $V'Y = \Lambda U'X + E$, where $E \sim N(0, I_r - \Lambda^2)$. Then (10) is a least square formulation of the regression interpretation. However, CCA is different from regression because the response $V'Y$ depends on an unknown V. Comparing (10) with (12), it is clear that (10) is a sparsity constrained version of (12) 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 taken in (11) to transform it to an estimator of U.

We now state a bound for the final estimator (11).

THEOREM 3.1. *Assume*

$$\frac{1}{n}\left(r(s_u + s_v) + s_u \log\frac{ep}{s_u} + s_v \log\frac{em}{s_v}\right) \le c \tag{13}$$

for some sufficiently small constant $c > 0$. Then there exist constants $C, C' > 0$ only depending on c such that

$$L(\widehat{U}, U) \le \frac{C}{n\lambda^2} s_u \left(r + \log\frac{ep}{s_u}\right), \tag{14}$$

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

REMARK 3.1. The paper assumes that M is a constant. However, the minimax upper bound of Theorem 3.1 does not depend on M even if M is allowed to grow with n. To be specific, assume the eigenvalues of Σ_x are bounded in the interval $[M_1, M_2]$. The convergence rate of $L(\widehat{U}, U)$ would still be $\frac{1}{n\lambda^2} s_u (r + \log\frac{ep}{s_u})$, because the dependence on M_1, M_2 has been implicitly built into the prediction loss. On the other hand, a convergence rate for the loss $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$ would be $(\frac{M_2}{M_1})\frac{1}{n\lambda^2} s_u (r + \log\frac{ep}{s_u})$, with an extra factor of the condition number of Σ_x.

Under assumption (13), Theorem 3.1 achieves a convergence rate for the prediction loss in U that does not depend on any parameter related to V, though the probability tail still involves m and s_v. However, it can be shown that $\exp(-C'(s_v + \log(em/s_v))) \le m^{-C'/2}$, and so the corresponding term in the tail probability goes to 0 as long as $m \to \infty$. The optimality of this upper bound can be justified by the following minimax lower bound.

THEOREM 3.2. *Assume that* $r \le \frac{s_u \wedge s_v}{2}$. *Then there exists some constant* $C > 0$ *only depending on* M *and an absolute constant* $c_0 > 0$, *such that*

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

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

By Theorems 3.1 and 3.2, the rate in (14), whenever it is upper bounded by a constant, is the minimax rate of the problem.

4. Adaptive and computationally efficient estimation. 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 it 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 two-stage convex program approach towards sparse CCA. The procedure is named CoLaR, standing for Convex program 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. The issues related to this 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). 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 $\mathrm{Tr}(L'\widehat{\Sigma}_{xy}R) = \langle \widehat{\Sigma}_{xy}, LR' \rangle$. 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 $|\mathrm{supp}(L)| \le s_u$, $|\mathrm{supp}(R)| \le s_v$ imply that the vector ℓ_0 norm $\|LR'\|_0 \le s_u s_v$. 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. The program (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

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

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

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

is the smallest convex set containing $\mathcal{O}_r$. The relation (17) is stated in the proof of Theorem 3 of [35]. Combining (15)–(17), we use the following convex program for the first stage in our adaptive estimation scheme:

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

Implementation of (18) is discussed in Section 10 in the supplement [17].

REMARK 4.1. A related but different convex relaxation was proposed in [32] 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' is suboptimal compared to the optimal rates determined in [16]; see Theorem 4.1 below. Nonetheless, it leads to a reasonable estimator for the subspaces spanned by the first r left and right canonical coefficient vectors under a sample size condition, which is sufficient for achieving the optimal estimation rates for U and V in a further refinement stage to be introduced below. Although it is possible that some other penalty function rather than the ℓ_1 penalty in (18) could also achieve this goal, ℓ_1 is appealing due to its simplicity.

Second stage. Now we turn to the convex relaxation to (10) in the second stage. By the discussion following Theorem 3.1, 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 [38], 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:

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

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. 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

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

As before, sample splitting is only used for technical arguments in the proof. Simulation results in Section 11 in the supplement [17] show that using the whole dataset repeatedly in (18)–(20) yields satisfactory performance and the improvement by the second stage is considerable.

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

$$\|\widehat{A} - UV'\|_{\mathrm{F}}^2 \leq C \frac{s_u s_v \log(p+m)}{n\lambda^2},$$

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

The error bound in Theorem 4.1 can be much larger than the optimal rate for joint estimation of UV' established in [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 $C_1 \geq 0$. Then there exist constants γ and γ_u only depending on C_1 and M such that if we set $\rho = \gamma'\sqrt{[\log(p+m)]/n}$ and $\rho_u = \gamma_u'\sqrt{(r+\log p)/n}$ 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 exist a constants $C, C' > 0$ only depending on 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 $\mathbb{P} \in \mathcal{P}(n, s_u, s_v, p, m, r, \lambda; M)$.

REMARK 4.3. The result of Theorem 4.2 assumes a constant M. Explicit dependence on the eigenvalues of the marginal covariance can be tracked even when M is diverging. Assuming the eigenvalues of Σ_x all lie in the interval $[M_1, M_2]$, then the convergence rate of $L(\widehat{U}, U)$ would be $(\frac{M_2}{M_1})^2 \frac{s_u(r+\log p)}{n\lambda^2}$ and a convergence rate of $\|P_{\widehat{U}} - P_U\|_{\mathrm{F}}^2$ would be $(\frac{M_2}{M_1})^3 \frac{s_u(r+\log p)}{n\lambda^2}$. Compared with Remark 3.1, there is an extra factor $(\frac{M_2}{M_1})^2$, which is also present for the Lasso error bounds [7, 28]. Evidence has been given in the literature that such an extra factor can be intrinsic to all polynomial-time algorithms [39].

Although both Theorems 4.1 and 4.2 assume Gaussian distributions, a scrutiny of the proofs shows that the same results hold if the Gaussian assumption is weakened to sub-Gaussian. By Theorem 3.2, the rate in Theorem 4.2 is optimal. By Theorems 4.1 and 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. A full treatment of adaptation to M is beyond the scope of the current paper, though it seems possible in view of the recent proposals in [5, 11, 31]. A careful examination of the proofs shows that the dependence of ρ and ρ_u on M is through $\|\Sigma_x\|_{\mathrm{op}}^{1/2}\|\Sigma_y\|_{\mathrm{op}}^{1/2}$ and $\|\Sigma_x\|_{\mathrm{op}}$, respectively. When p and m are bounded from above by a constant multiple of n, we can upper bound the operator norms by the sample counterparts to remove the dependence of these penalty parameters on M. We conclude this section with two more remarks.

REMARK 4.4. The group sparsity penalty used in the second stage (19) plays an important role in achieving the optimal rate $s_u(r + \log p)/(n\lambda^2)$. Except for the extra λ^{-2} term, this convergence rate is a typical one for group Lasso [24]. If we simply use an ℓ_1 penalty, then we will obtain the rate $rs_u \log p/(n\lambda^2)$, which is clearly sub-optimal.

REMARK 4.5. Comparing Theorem 3.1 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). We examine the necessity of this condition in more details in Section 5 below.

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 probably unavoidable for any computationally feasible estimator to be consistent. To be specific, we show that for a sequence of parameter spaces in (5)–(6), if the condition is violated, then any computationally efficient consistent estimator of sparse canonical coefficients leads to a computationally efficient and statistically powerful test for the Planted Clique detection problem in a regime where it is believed to be computationally intractable.

Planted clique. 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:

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

It is widely believed that when $k = O(N^{1/2-\delta})$, the problem (22) cannot be solved by any randomized polynomial-time algorithm. In the rest of the paper, we formalize the conjectured hardness of Planted Clique problem 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,*

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

Evidence supporting this hypothesis has been provided in [15, 30]. 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 [34] in classes defined by a restricted covariance concentration condition, submatrix detection [26] and community detection [18].

Necessity of the sample size condition (21). Under Hypothesis A, the necessity of condition (21) is supported by the following theorem.

THEOREM 5.1. *Suppose that Hypothesis A holds and that as $n \to \infty$, $p = m$ satisfying $2n \leq p \leq n^a$ for some constant $a > 1$, $s_u = s_v$, $n(\log n)^5 \leq c s_u^4$ for some sufficiently small $c > 0$, and $\lambda = \frac{s_u s_v}{7290 n (\log(12n))^2}$. If, for some $\delta \in (0, 1)$,*

$$\liminf_{n\to\infty} \frac{(s_u s_v)^{1-\delta} \log(p + m)}{n\lambda^2} > 0, \tag{23}$$

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

$$\liminf_{n\to\infty} \sup_{\mathbb{P} \in \mathcal{P}(n, s_u, s_v, p, m, 1, \lambda; 3)} \mathbb{P}\left\{ L(\widehat{u}, u) > \frac{1}{300} \right\} > \frac{1}{4}. \tag{24}$$

Comparing (21) with (23), we see that subject to a sub-polynomial factor, the condition (21) is necessary to achieve consistent sparse CCA estimation within polynomial time complexity.

REMARK 5.1. The statement in Theorem 5.1 is rigorous only if we assume the computational complexities of basic arithmetic operations on real numbers and sampling from univariate continuous distributions with analytic density functions are all $\Theta(1)$ [9]. To be rigorous under the probabilistic Turing machine model [2], we need to introduce appropriate discretization of the problem and be more careful with the complexity of random number generation. To convey the key ideas in our computational lower bound construction, we focus on the continuous case throughout this section and defer the formal discretization arguments to Section 8 in the supplement [17].

We divide the reduction argument leading to Theorem 5.1 into two parts. In the first part, we show Hypothesis A implies the computational hardness of the sparse PCA problem under the Gaussian spiked covariance model. In the second part, we show computational hardness of sparse PCA implies that of sparse CCA as stated in Theorem 5.1.

5.1. *Hardness of sparse PCA under Gaussian spiked covariance model.* Gaussian single spiked model [22] refers to the distribution $N_p(0, \Sigma)$ where $\Sigma = \tau\theta\theta' + I_p$. Here, θ is the eigenvector of unit length and $\tau > 0$ is the eigenvalue. Define the following Gaussian single spiked model parameter space for sparse PCA:

$$\begin{aligned}&\mathcal{Q}(n, s, p, \lambda)\\ &\quad= \big\{\mathcal{L}(W_1, \ldots, W_n) : W_i \overset{\text{i.i.d.}}{\sim} N_p\big(0, \tau\theta\theta' + I_p\big),\\ &\qquad \|\theta\|_0 \le s, \tau \in [\lambda, 3\lambda]\big\}.\end{aligned} \tag{25}$$

The minimax estimation rate for θ under the loss $\|P_{\widehat{\theta}} - P_\theta\|_{\mathrm{F}}^2$ is $\frac{\lambda+1}{n\lambda^2} s \log \frac{ep}{s}$; see, for instance, [12]. However, to achieve the above minimax rate via computationally efficient methods such as those proposed in [8, 12, 25], researchers have required the sample size to satisfy $n \ge C\frac{s^2 \log p}{\lambda^2}$ for some sufficiently large constant $C > 0$. Moreover, no computationally efficient estimator is known to achieve consistency when the sample size condition is violated. As a first step toward the establishment of Theorem 5.1, we show that Hypothesis A implies hardness of sparse PCA under Gaussian spiked covariance model (25) when $\liminf_{n\to\infty} \frac{s^{2-\delta} \log p}{n\lambda^2} > 0$ for some $\delta > 0$.

Previous computational lower bounds for sparse PCA in [6, 34] cannot be used here directly because they are only valid for parameter spaces defined via the restricted covariance concentration (RCC) condition. As pointed out in [34], such parameter spaces include (but are not limited to) all sub-Gaussian distributions with sparse leading eigenvectors and the covariance matrices need not be of the spiked form $\Sigma = \tau\theta\theta' + I_p$. Therefore, the Gaussian single spiked model parameter space defined in (25) only constitutes a small subset of such RCC parameter

spaces. The goal of the present subsection is to establish the computational lower bound for the Gaussian single spiked model directly.

Suppose we have an estimator $\widehat{\theta} = \widehat{\theta}(W_1, \ldots, W_n)$ of the leading sparse eigenvector, we propose the following reduction scheme to transform it into a test for (22). We first introduce some additional notation. Consider integers k and N. Define

$$\delta_N = \frac{k}{N}, \qquad \eta_N = \frac{k}{45N(\log N)^2}. \tag{26}$$

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

$$\bar{\phi}_\mu = \frac{1}{2}(\phi_\mu + \phi_{-\mu}) \tag{27}$$

denote the density function of the Gaussian mixture $\frac{1}{2}N(\mu, 1) + \frac{1}{2}N(-\mu, 1)$. 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 $|\mu| \le 3\sqrt{\eta_N \log N}$, define two probability distributions $\mathcal{F}_{\mu,0}$ and $\mathcal{F}_{\mu,1}$ with densities

$$f_{\mu,0}(x) = M_0\big(\phi_0(x) - \delta_N^{-1}\big[\bar{\phi}_\mu(x) - \phi_0(x)\big]\big)\mathbf{1}_{\{|x| \le 3\sqrt{\log N}\}}, \tag{28}$$

$$f_{\mu,1}(x) = M_1\big(\phi_0(x) + \delta_N^{-1}\big[\bar{\phi}_\mu(x) - \phi_0(x)\big]\big)\mathbf{1}_{\{|x| \le 3\sqrt{\log N}\}}, \tag{29}$$

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 $|\mu| \le 3\sqrt{\eta_N \log N}$. For details, see Lemma 7.4 in the supplement [17].

With the foregoing definition, the proposed reduction scheme can be summarized as Algorithm 1. Here, the starting point is the adjacency matrix A of the random graph, and the reduction is well defined for all instances of $N \ge 12n$ and $p \ge 2n$.

We now explain how the reduction achieves its goal. For simplicity, focus on the case where $p = 2n$. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{2n}) \in \{0, 1\}^{2n}$ where ε_i is the indicator of whether the ith row of A_0 (defined in Step 2 of Algorithm 1) belongs to the planted clique or not, and $\gamma = (\gamma_1, \ldots, \gamma_{2n})$ the indicators of the columns of A_0. We discuss the distributions of W 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 (cf., Lemma 7.1 in the supplement [17]). Hence, the rows of W are close to i.i.d. random vectors from the $N_p(0, I_p)$ distribution. Since $\widehat{\theta} = \widehat{\theta}(W_1, \ldots, W_n)$ is independent of $\{W_i\}_{i=n+1}^{2n}$, the LHS of (33) is close in distribution to a χ_n^2 random variable scaled by n which concentrates around its expected value one. Indeed, it is upper bounded by $1 + O(\sqrt{\log(n)/n})$ with high probability.

Algorithm 1: Reduction from planted clique to sparse PCA (in Gaussian single spiked model)

Input:
1. Graph adjacency matrix $A \in \{0,1\}^{N\times N}$;
2. Estimator $\widehat{\theta}$ for the leading eigenvector θ.

Output: A solution to the hypothesis testing problem (22).

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

$$\mu_i = \eta_N^{1/2}\xi_i, \qquad i = 1, \dots, 2n. \tag{30}$$

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

$$\mathcal{L}\big((B_0)_{ij}|\mu_i\big) = \mathcal{F}_{\mu_i,0} \quad \text{and} \quad \mathcal{L}\big((B_1)_{ij}|\mu_i\big) = \mathcal{F}_{\mu_i,1}. \tag{31}$$

Let $A_0 \in \{0,1\}^{2n\times 2n}$ be the lower–left $2n \times 2n$ submatrix of the matrix A. Generate a matrix $W = [W_1', \dots, W_{2n}']' \in \mathbb{R}^{2n\times p}$ where for each $i \in [2n]$, if $j \in [2n]$, we set

$$W_{ij} = (B_0)_{ij}\big(1 - (A_0)_{ij}\big) + (B_1)_{ij}(A_0)_{ij}. \tag{32}$$

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

Test Construction. Let $\widehat{\theta} = \widehat{\theta}(W_1, \dots, W_n)$ be the estimator of the leading eigenvector by treating $\{W_i\}_{i=1}^n$ as data. It is normalized to be a unit vector. We reject H_0^G if

$$\widehat{\theta}'\left(\frac{1}{n}\sum_{i=n+1}^{2n} W_i W_i'\right)\widehat{\theta} \ge 1 + \frac{1}{4}k\eta_N. \tag{33}$$

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 $\varepsilon_i = \gamma_j = 1$. Moreover, the joint distribution of $\{\varepsilon_1, \dots, \varepsilon_{2n}, \gamma_1, \dots, \gamma_{2n}\}$ is close to that of $4n$ i.i.d. Bernoulli random variables $\{\widetilde{\varepsilon}_1, \dots, \widetilde{\varepsilon}_{2n}, \widetilde{\gamma}_1, \dots, \widetilde{\gamma}_{2n}\}$ with success probability $\delta_N = k/N$. For simplicity, suppose that these indicators are indeed i.i.d. Bernoulli(δ_N) variables $\{\widetilde{\varepsilon}_1, \dots, \widetilde{\varepsilon}_{2n}, \widetilde{\gamma}_1, \dots, \widetilde{\gamma}_{2n}\}$. Then one can show that 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 $\widetilde{\varepsilon}_i$, μ_i and $(A_0)_{ij}$, is approximately $N(0,1)$. In contrast, conditioning on $\widetilde{\gamma}_j = 1$, for any $i \in [2n]$, the conditional distribution of $(W_{ij}|\widetilde{\gamma}_j = 1)$ is approximately $N(0, 1+\eta_N)$. 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

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

that is, a Gaussian spiked covariance model in (25). Here, the leading eigenvector θ has sparsity level $|\operatorname{supp}(\theta)| = |\operatorname{supp}(\widetilde{\gamma})| = \sum_j \widetilde{\gamma}_j$, which concentrates around its mean value $n\delta_N \asymp k$ if $N \asymp n$. Thus, if $\widehat{\theta}$ estimates θ well, then the LHS of (33) approximately follows a noncentral χ^2_n distribution scaled by n (with noncentrality parameter determined by k and n), which exceeds $1 + k\eta_N/4$ with high probability under the alternative hypothesis. Hence, Algorithm 1 is expected to yield a test with small error for the Planted Clique problem (22) when $\widehat{\theta}$ is a good estimator. Under the conditions of Theorem 5.2 below, $k\eta_N$ can be as small as $C\sqrt{\log n/n}$ for some constant $C > 0$.

The materialization of the foregoing discussion leads to the following result which demonstrates quantitatively that a decent estimator of the leading sparse eigenvector results in a good test [by applying the reduction (30)–(33)] for the Planted Clique detection problem (22).

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

$$\sup_{\mathbb{Q} \in \mathcal{Q}(n, 3k/2, p, k\eta_N/2)} \mathbb{Q}\left\{ \|P_{\widehat{\theta}} - P_\theta\|_{\mathrm{F}}^2 > \frac{1}{3} \right\} \le \beta, \tag{34}$$

the test ψ defined by (30)–(33) satisfies

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

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

If the estimator $\widehat{\theta}$ is uniformly consistent over $\mathcal{Q}(n, 3k/2, p, k\eta_N/2)$, then β is close to zero. Hence, the conclusion of Theorem 5.2 implies that for appropriate growing sequences of n, N and k, the testing error for (22) can be made smaller than any fixed nonzero probability. Further invoking Hypothesis A, we obtain the following computational lower bounds for sparse PCA.

THEOREM 5.3. *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 $\delta \in (0, 2)$,*

$$\liminf_{n \to \infty} \frac{s^{2-\delta} \log p}{n\lambda^2} > 0, \tag{35}$$

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

$$\liminf_{n \to \infty} \sup_{\mathbb{Q} \in \mathcal{Q}(n, s, p, \lambda)} \mathbb{Q}\left\{ \|P_{\widehat{\theta}} - P_\theta\|_{\mathrm{F}}^2 > \frac{1}{3} \right\} > \frac{1}{4}. \tag{36}$$

Algorithm 2: Reduction from sparse PCA to sparse CCA

Input:
1. Observations $W_1, \dots, W_n \in \mathbb{R}^p$;
2. Estimator $\widehat{u}$ of the first leading canonical correlation coefficient u.

Output: An estimator $\widehat{\theta}$ of the leading eigenvector of $\mathcal{L}(W_1)$.

1 Generate i.i.d. random vectors $Z_1, \dots, Z_n \sim N_p(0, I_p)$. Set

$$X_i = \frac{1}{\sqrt{2}}(W_i + Z_i), \qquad Y_i = \frac{1}{\sqrt{2}}(W_i - Z_i), \qquad i = 1, \dots, n. \tag{39}$$

2 Compute $\widehat{u} = \widehat{u}(X_1, Y_1, \dots, X_n, Y_n)$. Set

$$\widehat{\theta} = \widehat{\theta}(W_1, \dots, W_n) = \widehat{u}/\|\widehat{u}\|. \tag{40}$$

In addition to estimation, we can also consider the following sparse PCA detection problem: Let $\mathbb{Q}$ denote the joint distribution of $W_1, \dots, W_n$, and we want to test

$$H_0 : \mathbb{Q} \in \mathcal{Q}(n, s, p, 0) \quad \text{v.s.} \quad H_1 : \mathbb{Q} \in \mathcal{Q}(n, s, p, \lambda). \tag{37}$$

The space $\mathcal{Q}(n, s, p, 0)$ contains only one distribution $\mathbb{Q}_0$ where $W_i \overset{\text{i.i.d.}}{\sim} N_p(0, I_p)$. Given any testing procedure $\phi = \phi(W_1, \dots, W_n)$, we can obtain a solution to (22) by replacing the third step in Algorithm 1 with the direct testing result of $\phi(W_1, \dots, W_n)$. Following the lines of the proof of Theorem 5.3, we have the following theorem.

THEOREM 5.4. *Under the same condition of Theorem 5.3, for any randomized polynomial-time test ϕ for testing* (37),

$$\liminf_{n\to\infty}\Big(\mathbb{Q}_0\phi + \sup_{\mathbb{Q}\in\mathcal{Q}(n,s,p,\lambda)} \mathbb{Q}(1-\phi)\Big) \geq \frac{1}{4}. \tag{38}$$

REMARK 5.2. Theorems 5.3–5.4 are the first computational lower bounds for sparse PCA that are valid in the setting of Gaussian single spiked covariance models (25).

5.2. *Hardness of sparse CCA.* In the second step, we show that computational hardness of sparse PCA under Gaussian spiked covariance model implies the desired result in Theorem 5.1. We propose the reduction in Algorithm 2.

To see why Algorithm 2 is effective, one can verify that if $W_i \overset{\text{i.i.d.}}{\sim} N_p(0, \tau\theta\theta' + I_p)$, then $(X_i', Y_i')' \overset{\text{i.i.d.}}{\sim} N_{p+m}(0, \Sigma)$ where

$$\Sigma_x = \Sigma_y = \frac{\tau}{2}\theta\theta' + I_p, \qquad \Sigma_{xy} = \Sigma_x(\lambda uv')\Sigma_y \tag{41}$$

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). Thus, the leading eigenvector of W_i aligns with the leading canonical coefficient vectors of (X_i, Y_i). Exploiting this connection, we obtain the following theorem.

THEOREM 5.5. *Consider $p = m$, $s_u = s_v$ and $\lambda \le 1$. Then for any $\widehat{u}$ such that*

$$\sup_{\mathbb{P}\in\mathcal{P}(n,s_u,s_v,p,m,1,\lambda/3;3)} \mathbb{P}\left\{L(\widehat{u},u) > \frac{1}{300}\right\} \le \beta, \tag{42}$$

the estimator $\widehat{\theta}$ defined by Algorithm 2 satisfies

$$\sup_{\mathbb{Q}\in\mathcal{Q}(n,s,p,\lambda)} \mathbb{Q}\left\{\|P_{\widehat{\theta}} - P_\theta\|_{\mathrm{F}}^2 > \frac{1}{3}\right\} \le \beta.$$

If we start with an estimator $\widehat{u}$ of the leading canonical coefficient vector, then we can construct the reduction from Planted Clique to sparse CCA directly by essentially following the steps in Algorithm 1 while using Algorithm 2 to construct $\widehat{\theta}$ from $\widehat{u}$ in the third step. Finally, the desired Theorem 5.1 is a direct consequence of Theorems 5.3 and 5.5.

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 [17].

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 9.3 in the supplement [17]. First, observe 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

$$\varepsilon_{n,u} = \sqrt{\frac{1}{n}\left(s_u + \log\frac{ep}{s_u}\right)}, \qquad \varepsilon_{n,v} = \sqrt{\frac{1}{n}\left(s_v + \log\frac{em}{s_v}\right)}. \tag{43}$$

The following lemma bounds the normalization effect.

LEMMA 6.1. *Assume $\varepsilon_{n,u}^2 + \varepsilon_{n,v}^2 \le c$ for some sufficiently small constant $c \in (0,1)$. Then there exist some constants $C, C' > 0$ only depending on c such that*

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

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

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*

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

As was argued in Section 4.1, the set $\mathcal{C}_r$ is the convex hull of $\mathcal{O}_r$. The following curvature lemma shows that the relaxation $\mathcal{C}_r$ preserves the restricted strong convexity of the objective function.

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

$$\langle FKG', FG' - E\rangle \geq \frac{d_r}{2}\|FG' - E\|_{\mathrm{F}}^2 - \|K - D\|_{\mathrm{F}}\|FG' - E\|_{\mathrm{F}}. \tag{44}$$

Define

$$\widetilde{\Sigma}_{xy} = \widehat{\Sigma}_x^{(0)} U \Lambda V' \widehat{\Sigma}_y^{(0)}. \tag{45}$$

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

LEMMA 6.4. *Assume $r\sqrt{[\log(p+m)]/n} \leq c$ for some sufficiently small constant $c \in (0,1)$. Then there exist some constants $C, C' > 0$ only depending on M and c such that $\|\widehat{\Sigma}_{xy}^{(0)} - \widetilde{\Sigma}_{xy}\|_\infty \leq C\sqrt{[\log(p+m)]/n}$, with probability at least $1 - (p+m)^{-C'}$.*

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

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

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

LEMMA 6.5. *Assume $\frac{1}{n}((k_u \wedge p)\log(ep/(k_u \wedge p)) + (k_v \wedge m)\log(em/(k_v \wedge m))) \leq c$ for some sufficiently small constant $c > 0$. Then there exist some constants $C, C' > 0$ only depending on M and c such that for $\delta_u(k_u) = \sqrt{\frac{(k_u\wedge p)\log(ep/(k_u\wedge p))}{n}}$ and $\delta_v(k_v) = \sqrt{\frac{(k_v\wedge m)\log(em/(k_v\wedge m))}{n}}$, we have*

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

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

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

Proofs of Lemmas 6.1–6.5 are given in Section 9.3.1 of the supplement [17].

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 $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$. In the second step, we derive a generalized cone condition and use it to lower bound $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$ by a constant multiple of $\|\Delta\|_{\mathrm{F}}$ and hence the upper bound on $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$ leads to an upper bound on $\|\Delta\|_{\mathrm{F}}$.

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

$$\|\Sigma_x^{1/2}(\widetilde{A} - UV')\Sigma_y^{1/2}\|_{\mathrm{op}} \le C(\varepsilon_{n,u} + \varepsilon_{n,v}). \tag{46}$$

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

$$-\langle\widetilde{\Sigma}_{xy}, \Delta\rangle \le \rho\big(\|\widetilde{A}\|_1 - \|\widetilde{A} + \Delta\|_1\big) + \langle\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}, \Delta\rangle, \tag{47}$$

where $\widetilde{\Sigma}_{xy}$ is defined in (45). For the first term on the RHS of (47), we have

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

For the second term on the RHS of (47), we have $\langle\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}, \Delta\rangle \le \|\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}\|_\infty\|\Delta\|_1$. Thus, when

$$\rho \ge 2\|\widehat{\Sigma}_{xy} - \widetilde{\Sigma}_{xy}\|_\infty, \tag{48}$$

we have

$$-\langle\widetilde{\Sigma}_{xy}, \Delta\rangle \le \frac{3\rho}{2}\|\Delta_{S_uS_v}\|_1 - \frac{\rho}{2}\|\Delta_{(S_uS_v)^c}\|_1. \tag{49}$$

Using Lemma 6.3, we can lower bound the LHS of (49) as

$$\begin{aligned}-\langle\widetilde{\Sigma}_{xy}, \Delta\rangle &= \langle\widehat{\Sigma}_x^{1/2}U\Lambda V'\widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A})\widehat{\Sigma}_y^{1/2}\rangle\\ &= \langle\widehat{\Sigma}_x^{1/2}\widetilde{U}\widetilde{\Lambda}\widetilde{V}'\widehat{\Sigma}_y^{1/2}, \widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A})\widehat{\Sigma}_y^{1/2}\rangle \qquad (50)\\ &\ge \frac{1}{2}\lambda_r\|\widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A})\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 - \delta\|\widehat{\Sigma}_x^{1/2}(\widetilde{A} - \widehat{A})\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}},\end{aligned}$$

where $\delta = \|\widetilde{\Lambda} - \Lambda\|_{\mathrm{F}}$. Combining (49) and (50), we have

$$\lambda_r\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 \le 3\rho\|\Delta_{S_uS_v}\|_1 - \rho\|\Delta_{(S_uS_v)^c}\|_1 + 2\delta\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}} \tag{51}$$

$$\le 3\rho\|\Delta_{S_uS_v}\|_1 + 2\delta\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}. \tag{52}$$

Solving the quadratic equation (52) by Lemma 2 of [12], we have

$$\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 \le 6\rho\|\Delta_{S_uS_v}\|_1/\lambda_r + 4\delta^2/\lambda_r^2. \tag{53}$$

Combining (51) and (53), we have

$$\begin{aligned} 0 &\le 3\rho\|\Delta_{S_uS_v}\|_1 - \rho\|\Delta_{(S_uS_v)^c}\|_1 + \delta^2/\lambda_r + \lambda_r\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 \\ &\le 9\rho\|\Delta_{S_uS_v}\|_1 - \rho\|\Delta_{(S_uS_v)^c}\|_1 + 5\delta^2/\lambda_r, \end{aligned} \tag{54}$$

which gives rise to the generalized cone condition that we are going to use in Step 2. Finally, by the bound $\|\Delta_{S_uS_v}\|_1 \le \sqrt{s_us_v}\rho\|\Delta_{S_uS_v}\|_{\mathrm{F}}$ and (53), we have

$$\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}^2 \le 6\sqrt{s_us_v}\rho\|\Delta_{S_uS_v}\|_{\mathrm{F}}/\lambda_r + 4\delta^2/\lambda_r^2, \tag{55}$$

which completes the first step.

Step 2. By (54), we have obtained the following condition:

$$\|\Delta_{(S_uS_v)^c}\|_1 \le 9\|\Delta_{S_uS_v}\|_1 + 5\delta^2/(\rho\lambda_r). \tag{56}$$

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 $\|\widehat{\Sigma}_x^{1/2}\Delta\widehat{\Sigma}_y^{1/2}\|_{\mathrm{F}}$ by $\|\Delta\|_{\mathrm{F}}$ on the generalized cone. Motivated by the argument in [7], 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 $J_2, \ldots, J_K$ of size t (with $|J_K| \le t$), such that J_k is the set of (double) indices corresponding to the entries of t largest absolute values in Δ outside $\widetilde{J} \cup \bigcup_{j=2}^{k-1} J_j$. By the triangle inequality,

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

By the construction of J_k, we have

$$\begin{aligned} \sum_{k=2}^K \|\Delta_{J_k}\|_{\mathrm{F}} &\le \sqrt{t}\sum_{k=2}^K \|\Delta_{J_k}\|_\infty \\ &\le t^{-1/2}\sum_{k=2}^K \|\Delta_{J_{k-1}}\|_1 \le t^{-1/2}\|\Delta_{(S_uS_v)^c}\|_1 \\ &\le t^{-1/2}\left(9\|\Delta_{S_uS_v}\|_1 + \frac{5\delta^2}{\rho\lambda_r}\right) \le 9\sqrt{\frac{s_us_v}{t}}\|\Delta_{\widetilde{J}}\|_{\mathrm{F}} + \frac{5\delta^2}{\rho\lambda_r\sqrt{t}}, \end{aligned} \tag{57}$$

where we have used the generalized cone condition (56). Hence, we have the lower bound

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

with

$$\begin{aligned}\kappa_1 &= \sqrt{\phi_{\min}^{\widehat{\Sigma}_x}(s_u+t)\phi_{\min}^{\widehat{\Sigma}_y}(s_v+t)} - 9\sqrt{\frac{s_u s_v}{t}}\sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)\phi_{\max}^{\widehat{\Sigma}_y}(t)},\\ \kappa_2 &= 5\sqrt{\phi_{\max}^{\widehat{\Sigma}_x}(t)\phi_{\max}^{\widehat{\Sigma}_y}(t)}.\end{aligned} \tag{58}$$

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, observe that by Lemma 6.5, (58) 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 $9c_1^{-1/2}\sqrt{M + C\delta_u(c_1 s_u s_v)} \times \sqrt{M + C\delta_v(c_1 s_u s_v)}$, 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 (13). 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 (13). The same argument also holds for the other terms. Similarly, κ_2 can be upper bounded by some constant.

Together with (55), this brings the inequality

$$\|\Delta_{\tilde{J}}\|_{\mathrm{F}}^2 \leq C_1(\sqrt{s_u s_v}\rho/\lambda_r)\|\Delta_{\tilde{J}}\|_{\mathrm{F}} + C_2\big(\delta^2/\lambda_r^2 + \big(\delta^2/(\rho\lambda_r\sqrt{t})\big)^2\big).$$

Solving this quadratic equation, we have

$$\|\Delta_{\tilde{J}}\|_{\mathrm{F}}^2 \leq C\bigg(\frac{s_u s_v\rho^2}{\lambda_r^2} + \frac{\delta^2}{\lambda_r^2} + \bigg(\frac{\delta^2}{\rho\lambda_r\sqrt{t}}\bigg)^2\bigg). \tag{59}$$

By (57), we have

$$\|\Delta_{\tilde{J}^c}\|_{\mathrm{F}} \leq \sum_{k=2}^{K}\|\Delta_{J_k}\|_{\mathrm{F}} \leq 9\sqrt{\frac{s_u s_v}{t}}\|\Delta_{\tilde{J}}\|_{\mathrm{F}} + \frac{5\delta^2}{\rho\lambda_r\sqrt{t}}. \tag{60}$$

Summing (59) and (60), we obtain a bound for $\|\Delta\|_{\mathrm{F}}$. According to Lemma 6.4, we may choose $\rho = \gamma\sqrt{[\log(p+m)]/n}$ for some large γ, so that (48) holds with high probability. By Lemma 6.1, $\delta \leq C\sqrt{r(s_u + s_v + \log(p+m))/n} \leq C'\rho\sqrt{t}$ with high probability. Hence,

$$\|\Delta\|_{\mathrm{F}} \leq C\sqrt{s_u s_v}\rho/\lambda_r, \tag{61}$$

with high probability. This completes the second step. Finally, the triangle inequality leads to $\|\widehat{A} - UV'\|_{\mathrm{F}} \leq \|\Delta\|_{\mathrm{F}} + \|\widetilde{A} - UV'\|_{\mathrm{F}}$. By (46) and (61), the proof is complete. □

6.2. *Proof of Theorem 4.2.* Define $U^* = U\Lambda V'\Sigma_y \widehat{V}^{(0)}$ and $\Delta = \widehat{U}^{(1)} - U^*$.

LEMMA 6.6. *Assume* $\frac{r+\log p}{n} \le c$ *for some sufficiently small constant* $c \in (0,1)$. *Then there exist some constants* $C, C' > 0$ *only depending on* M *and* c *such that* $\max_{1\le j\le p} \|[\widehat{\Sigma}_{xy}^{(1)}\widehat{V}^{(0)} - \widehat{\Sigma}_x^{(1)}U^*]_{j\cdot}\| \le C\sqrt{(r+\log p)/n}$, *with probability at least* $1 - \exp(-C'(r+\log p))$.

The proof of Lemma 6.6 is given in Section 9.3.1 of the supplement [17].

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. These covariance matrices depend 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 $\mathsf{Tr}((\widehat{U}^{(1)})'\widehat{\Sigma}_x\widehat{U}^{(1)}) - 2\mathsf{Tr}((\widehat{U}^{(1)})' \times \widehat{\Sigma}_{xy}\widehat{V}^{(0)}) + \rho_u\sum_{j=1}^p \|\widehat{U}_{j\cdot}^{(1)}\| \le \mathsf{Tr}((U^*)'\widehat{\Sigma}_x U^*) - 2\mathsf{Tr}((U^*)'\widehat{\Sigma}_{xy}\widehat{V}^{(0)}) + \rho_u\sum_{j=1}^p \|U_{j\cdot}^*\|$. After rearrangement, we have

$$\mathsf{Tr}\big(\Delta'\widehat{\Sigma}_x\Delta\big) \le \rho_u \sum_{j=1}^p \big[\|U_{j\cdot}^*\| - \|U_{j\cdot}^* + \Delta_{j\cdot}\|\big] + 2\,\mathsf{Tr}\big[\Delta'\big(\widehat{\Sigma}_{xy}\widehat{V}^{(0)} - \widehat{\Sigma}_x U^*\big)\big]. \tag{62}$$

For the first term on the RHS of (62), we have

$$\begin{aligned}\sum_{j=1}^p \big(\|U_{j\cdot}^*\| - \|U_{j\cdot}^* + \Delta_{j\cdot}\|\big) &= \sum_{j\in S_u}\|U_{j\cdot}^*\| - \sum_{j\in S_u}\|U_{j\cdot}^* + \Delta_{j\cdot}\| - \sum_{j\in S_u^c}\|\Delta_{j\cdot}\| \\ &\le \sum_{j\in S_u}\|\Delta_{j\cdot}\| - \sum_{j\in S_u^c}\|\Delta_{j\cdot}\|.\end{aligned}$$

For the second term on the RHS of (62), we have

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

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

$$\rho_u \ge 4 \max_{1\le j\le p} \big\|\big[\widehat{\Sigma}_{xy}\widehat{V}^{(0)} - \widehat{\Sigma}_x U^*\big]_{j\cdot}\big\|, \tag{63}$$

we have

$$\mathsf{Tr}\big(\Delta'\widehat{\Sigma}_x\Delta\big) \le \frac{3\rho_u}{2}\sum_{j\in S_u}\|\Delta_{j\cdot}\| - \frac{\rho_u}{2}\sum_{j\in S_u^c}\|\Delta_{j\cdot}\|. \tag{64}$$

Since $\sum_{j\in S_u}\|\Delta_{j\cdot}\| \le \sqrt{s_u}\sqrt{\sum_{j\in S_u}\|\Delta_{j\cdot}\|^2}$, (64) can be upper bounded by

$$\mathsf{Tr}(\Delta'\widehat{\Sigma}_x\Delta) \le \frac{3\sqrt{s_u}\rho_u}{2}\sqrt{\sum_{j\in S_u}\|\Delta_{j\cdot}\|^2}. \tag{65}$$

This completes the first step.

Step 2. The inequality (64) implies the cone condition

$$\sum_{j\in S_u^c}\|\Delta_{j\cdot}\| \le 3\sum_{j\in S_u}\|\Delta_{j\cdot}\|. \tag{66}$$

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

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

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

$$\begin{aligned}\sum_{k\ge 2}\|\Delta_{J_k *}\|_{\mathrm{F}} &\le \sqrt{t}\sum_{k\ge 2}\max_{j\in J_k}\|\Delta_{j\cdot}\| \le \sqrt{t}\sum_{k\ge 2}\frac{1}{t}\sum_{j\in J_{k-1}}\|\Delta_{j\cdot}\| \\ &\le t^{-1/2}\sum_{j\in S_u^c}\|\Delta_{j\cdot}\| \le 3t^{-1/2}\sum_{j\in S_u}\|\Delta_{j\cdot}\| \qquad (67)\\ &\le 3\sqrt{\frac{s_u}{t}}\sqrt{\sum_{j\in S_u}\|\Delta_{j\cdot}\|^2} \le 3\sqrt{\frac{s_u}{t}}\|\Delta_{\widetilde{S}_u *}\|_{\mathrm{F}}. \qquad (68)\end{aligned}$$

In the above derivation, we have used the construction of J_k and the cone condition (66). Hence, $\|n^{-1/2}X\Delta\|_{\mathrm{F}} \ge \kappa\|\Delta_{\widetilde{S}_u *}\|_{\mathrm{F}}$ 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 (65), we have

$$\|\Delta_{\widetilde{S}_u *}\|_{\mathrm{F}} \le C\sqrt{s_u}\rho_u/(2\kappa_0^2). \tag{69}$$

By (67)–(68), we have

$$\|\Delta_{(\widetilde{S}_u)^c *}\|_{\mathrm{F}} \le \sum_{k\ge 2}\|\Delta_{J_k *}\|_{\mathrm{F}} \le 3\sqrt{s_u/t}\|\Delta_{\widetilde{S}_u *}\|_{\mathrm{F}} \le 3c_1^{-1/2}\|\Delta_{\widetilde{S}_u *}\|_{\mathrm{F}}. \tag{70}$$

Summing (69) and (70), we have $\|\Delta\|_{\mathrm{F}} \le C\sqrt{s_u}\rho$. By Lemma 6.6, we may choose $\rho_u \ge \gamma_u\sqrt{\frac{r+\log p}{n}}$ for some large γ_u so that (63) holds with high probability. Hence, $\|\Delta\|_{\mathrm{F}} \le C\sqrt{s_u(r+\log p)/n}$ with high probability. This completes the second step.

Step 3. Using the same argument in Step 2 of the proof of Theorem 3.1 (see supplementary material [17]), we obtain the desired bound for $L(\widehat{U}, U)$. The proof is complete. □

SUPPLEMENTARY MATERIAL

Supplement to "Sparse CCA: Adaptive estimation and computational barriers" (DOI: 10.1214/16-AOS1519SUPP; .pdf). The supplement presents additional proofs and technical details, implementation detail of (18), and numerical studies.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] ARORA, S. and BARAK, B. (2009). *Computational Complexity*: *A Modern Approach*. Cambridge Univ. Press, Cambridge. MR2500087

[3] AVANTS, B. B., COOK, P. A., UNGAR, L., GEE, J. C. and GROSSMAN, M. (2010). Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage* **50** 1004–1016.

[4] BAO, Z., HU, J., PAN, G. and ZHOU, W. (2014). Canonical correlation coefficients of high-dimensional normal vectors: Finite rank case. Preprint. Available at arXiv:1407.7194.

[5] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[6] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066.

[7] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[8] BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

[9] BLUM, L., CUCKER, F., SHUB, M. and SMALE, S. (2012). *Complexity and Real Computation*. Springer, Berlin.

[10] BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Faund. Trends Mach. Learn.* **3** 1–122.

[11] BUNEA, F., LEDERER, J. and SHE, Y. (2014). The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inform. Theory* **60** 1313–1325.

[12] CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110.

[13] CHEN, M., GAO, C., REN, Z. and ZHOU, H. H. (2013). Sparse CCA via precision adjusted iterative thresholding. Preprint. Available at arXiv:1311.6186.

[14] DOUGLAS, J. JR. and RACHFORD, H. H. JR. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.* **82** 421–439. MR0084194

[15] FELDMAN, V., GRIGORESCU, E., REYZIN, L., VEMPALA, S. S. and XIAO, Y. (2013). Statistical algorithms and a lower bound for detecting planted cliques. In *STOC'*13—*Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 655–664. ACM, New York. MR3210827

[16] GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982

[17] GAO, C., MA, Z. and ZHOU, H. H. (2017). Supplement to "Sparse CCA: Adaptive Estimation and Computational Barriers." DOI:10.1214/16-AOS1519SUPP.

[18] HAJEK, B., WU, Y. and XU, J. (2014). Computational lower bounds for community detection on random graphs. Preprint. Available at arXiv:1406.6625.

[19] HARDOON, D. R. and SHAWE-TAYLOR, J. (2011). Sparse canonical correlation analysis. *Mach. Learn.* **83** 331–353. MR3108214

[20] HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

[21] JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010

[22] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

[23] LÊ CAO, K.-A., MARTIN, P. G. P., ROBERT-GRANIÉ, C. and BESSE, P. (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics* **10** 34.

[24] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

[25] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

[26] MA, Z. and WU, Y. (2013). Computational barriers in minimax submatrix detection. Preprint. Available at arXiv:1309.5914.

[27] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

[28] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

[29] PARKHOMENKO, E., TRITCHLER, D. and BEYENE, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.* **8** Art. 1, 36. MR2471148

[30] ROSSMAN, B. (2010). Average-case complexity of detecting cliques. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

[31] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[32] VU, V. Q., CHO, J., LEI, J. and ROHE, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems* 2670–2678.

[33] WAAIJENBORG, S. and ZWINDERMAN, A. H. (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics* **10** 315.

[34] WANG, T., BERTHET, Q. and SAMWORTH, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. Preprint. Available at arXiv:1408.5369.

[35] WATSON, G. A. (1993). On matrix approximation problems with Ky Fan k norms. *Numer. Algorithms* **5** 263–272.

[36] WIESEL, A., KLIGER, M. and HERO, A. O., III (2008). A greedy approach to sparse canonical correlation analysis. Preprint. Available at arXiv:0801.2748.

[37] WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

[38] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67.

[39] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. Preprint. Available at arxiv:1402.1918.

C. GAO
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
CHICAGO, ILLINOIS 60637
USA
E-MAIL: chaogao@galton.uchicago.edu
URL: http://www.stat.uchicago.edu/~chaogao

Z. MA
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: zongming@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~zongming

H. H. ZHOU
DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: huibin.zhou@yale.edu
URL: http://www.stat.yale.edu/~hz68