On Sparse Canonical Correlation Analysis

Yongchun Li, Santanu S. Dey, and Weijun Xie

School of Industrial and Systems Engineering, Georgia Institute of Technology

Abstract. The classical Canonical Correlation Analysis (CCA) identifies the correlations between two sets of multivariate variables based on their covariance, which has been widely applied in diverse fields such as computer vision, natural language processing, and speech analysis. Despite its popularity, CCA can encounter challenges in explaining correlations between two variable sets within high-dimensional data contexts. Thus, this paper studies Sparse Canonical Correlation Analysis (SCCA) that enhances the interpretability of CCA. We first show that SCCA generalizes three well-known sparse optimization problems, sparse PCA, sparse SVD, and sparse regression, which are all classified as NP-hard problems. This result motivates us to develop strong formulations and efficient algorithms. Our main contributions include (i) the introduction of a combinatorial formulation that captures the essence of SCCA and allows the development of approximation algorithms; (ii) the derivation of an equivalent mixed-integer semidefinite programming model that facilitates a specialized branch-and-cut algorithm with analytical cuts; and (iii) the establishment of the complexity results for two low-rank special cases of SCCA. The effectiveness of our proposed formulations and algorithms is validated through numerical experiments.

1 Introduction

The Canonical Correlation Analysis (CCA), proposed by H. Hotelling [18], aims to identify the correlations between two sets of multivariate variables based on their covariance. Since then, CCA has become a powerful statistical technique used for multivariate data analysis, with its applications across diverse fields such as computer vision [19], natural language processing [32], and speech analysis [16]. Despite its popularity, CCA can encounter challenges in explaining correlations between two variable sets within high-dimensional data contexts, such as genomic datasets [30]. In contrast, Sparse Canonical Correlation Analysis (SCCA), which seeks sparse linear combinations of these variable sets, offers substantially enhanced interpretability [35, 36, 38].

Formally, this paper studies the SCCA problem:

$$v^* := \max_{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m} \left\{ \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C} \boldsymbol{y} \leq 1, \|\boldsymbol{x}\|_0 \leq s_1, \|\boldsymbol{y}\|_0 \leq s_2 \right\},$$
(SCCA)

where $s_1 \leq n$, $s_2 \leq m$ are positive integers and $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{C} \end{pmatrix}$ denotes a covariance matrix of $(n + m)$ random variables. Specifically, $\boldsymbol{B}$ and $\boldsymbol{C}$ are the covariance

matrices of the n and m random variables, respectively, and $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ is the cross-covariance matrix between n and m random variables. Hence, $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{C} \end{pmatrix}, \boldsymbol{B}, \boldsymbol{C}$ are positive semidefinite matrices of size $(n+m)$, n, and m, respectively. Here, matrices $\boldsymbol{B}, \boldsymbol{C}$ can be singular, i.e., some random variables may be dependent on others. In fact, the covariance matrices $\boldsymbol{B}, \boldsymbol{C}$ are often low-rank, especially within the high-dimension low-sample size data context (see, e.g., the gene expression data in [35]).

The SCCA problem generalizes three widely-studied sparsity-constrained optimization problems as special cases, which are sparse PCA [2, 10, 22], sparse SVD [23, 35], and sparse regression [3, 17]. To be specific, when $n = m$, $s_1 = s_2$, $\boldsymbol{B}, \boldsymbol{C}$ are identity matrices, and $\boldsymbol{A}$ is a positive semidefinite matrix, SCCA reduces to the classic sparse PCA problem; when $\boldsymbol{B}, \boldsymbol{C}$ are identity matrices, SCCA becomes the sparse SVD problem; and when $\boldsymbol{A}$ is rank-one, Section 4 shows that SCCA is equivalent to two sparse linear regression subproblems.

1.1   Main contributions

SCCA is generally NP-hard, given that its special cases, sparse PCA, sparse SVD, and sparse regression are all classified as NP-hard problems. We are motivated to develop efficient formulations and algorithms for SCCA through a mixed-integer optimization lens. The main contributions, along with the structure of the remainder of this paper, are the following:

(i) In Section 2, we present an exact semidefinite programming (SDP) reformulation and derive a closed-form optimal value of classic CCA problem. We also develop an equivalent combinatorial formulation of SCCA;

(ii) Section 3 derives an equivalent mixed-integer SDP (MISDP) reformulation for SCCA. When applying the Benders decomposition approach, instead of solving the large-scale SDPs, we design a customized branch-and-cut algorithm with closed-form cuts, which can successfully solve SCCA to optimality;

(iii) When the covariance matrix $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{C} \end{pmatrix}$ is low-rank, Section 4 studies the complexity of two special cases of SCCA; and

(iv) Section 5 numerically test the proposed formulations and algorithms.

1.2   Relevant literature

SCCA. To the best of our knowledge, the work [30] was the first paper that introduced the concept of SCCA to select only small subsets of variables to better explain the relationship between many genetic loci and gene expression phenotypes. A handful subset of features enhances interpretability, a desirable property, especially in complex data analysis, which has been successfully demonstrated in Sparse PCA [20]. To obtain sparse canonical loadings $(\boldsymbol{x}, \boldsymbol{y})$, [33] first applied elastic net penalty to the classical CCA via an iterative regression procedure. In a seminal work on SCCA [35], the authors proposed a rigorous formulation by enforcing the ℓ_1 constraints on variables $(\boldsymbol{x}, \boldsymbol{y})$ and developed a penalized matrix decomposition method to solve the penalized CCA problem. Then, extensive research has

focused on various penalty norm functions to obtain sparse canonical loadings (see, e.g., [7, 15, 21, 33, 36]). In particular, [7] penalized multiple canonical loadings by ℓ_1 norm and computed the sparse solution by the linearized Bregman method. It should be noted that under the assumption that the leading canonical loadings are sparse, [5, 13, 14] established theoretical guarantees of iterative approaches for estimating sparse solutions. Another research direction in SCCA introduced penalty functions based on group structural information of input data and developed group SCCA methods [24, 25]. For a comprehensive overview of CCA and SCCA methods, we refer readers to the survey by [38] and the references therein. These approaches, however, do not strictly enforce the exact sparsity requirement but only approximate the sparsity requirement (i.e., the ℓ_0 norm) by a convex function. Another relevant work [34] introduced binary variables to recast SCCA as a mixed-integer nonconvex program under the assumption of positive definite matrices $\boldsymbol{B}, \boldsymbol{C}$, based on which they designed a branch-and-bound algorithm. Different from the literature, our work does not require positive definiteness assumption of matrices $\boldsymbol{B}, \boldsymbol{C}$ and we are able to obtain mixed-integer conic and semidefinite programming reformulations, allowing for better exact and approximation algorithms.

Connections to and differences with sparse PCA and sparse SVD. Analogous to SCCA, both sparse PCA [10, 20] and sparse SVD [23] select small subsets of variables to improve the interpretability of dimensionality reduction methods: PCA and SVD. Considerable investigation has been conducted on solving sparse PCA and sparse SVD from three angles: convex relaxations [9–11], approximation algorithms [4, 6, 23], and exact algorithms [2, 22, 23]. As mentioned before, in sparse PCA and sparse SVD, the covariance matrices $\boldsymbol{B}, \boldsymbol{C}$ are identity. Such a setting dramatically simplifies the subset selection problems of sparse PCA and sparse SVD compared to that of SCCA, as in these problems, it suffices to focus on the selection of a submatrix of the matrix $\boldsymbol{A}$. Specifically, it is shown in [8, 22, 29] that sparse PCA reduces to selecting a principal submatrix of $\boldsymbol{A}$ to maximize the largest eigenvalue(s) and sparse SVD reduces to selecting a possibly non-symmetric submatrix of $\boldsymbol{A}$ to maximize the largest singular value(s) [23]. Quite differently, the combinatorial reformulation (1) of SCCA aims to simultaneously select a sized-$(s_1 \times s_1)$ principal submatrix of $\boldsymbol{B}$, a sized-$(s_2 \times s_2)$ principal submatrix of $\boldsymbol{C}$, and a sized-$(s_1 \times s_2)$ submatrix of $\boldsymbol{A}$. These fundamental differences in the underlying formulations of sparse PCA and sparse SVD preclude the direct application of their existing algorithms to the SCCA.

Notations: The following notation is used throughout the paper. We use bold lower-case letters (e.g., $\boldsymbol{x}$) and bold upper-case letters (e.g., $\boldsymbol{X}$) to denote vectors and matrices, respectively, and we use corresponding non-bold letters (e.g., x_i) to denote their components. We let $\mathcal{S}^n, \mathcal{S}^n_+, \mathcal{S}^n_{++}$ denote the set of all the $n \times n$ symmetric real matrices, the set of all the $n \times n$ symmetric positive semidefinite matrices, and the set of all the $n \times n$ symmetric positive definite matrices, respectively. We let $\boldsymbol{I}$ denote the identity matrix and let $\boldsymbol{0}$ denote the vector or matrix with all-zero entries. We let $\mathbb{R}^n_+$ denote the set of all n-dimensional nonnegative vectors. We let $[n] := \{1, 2, \cdots, n\}, [s, n] := \{s, s + 1, \cdots, n\}$. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ and two subsets $S \subseteq [n], T \subseteq [m]$, we let $\boldsymbol{A}^\dagger$ denote the pseudo inverse

of matrix $\boldsymbol{A}$, let $\boldsymbol{A}_{S,T}$ denote a submatrix of $\boldsymbol{A}$ with rows and columns indexed by sets S, T, respectively, and let $(\boldsymbol{A}_{S,T})^{\dagger}$ denote the pseudo inverse of submatrix $\boldsymbol{A}_{S,T}$. For a set S and an integer k, we define the set $S + k := \{i + k | i \in S\}$. Given a vector $\boldsymbol{a} \in \mathbb{R}^n$ and a subset $S \subseteq [n]$, we let $\boldsymbol{a}_S$ denote a subvector of $\boldsymbol{a}$ in the subset S. We define $[\lambda]_{+} := \max\{\lambda, 0\}$. We let $\sigma_{\max}(\cdot)$ denote the largest singular value function and let $\lambda_{\max}(\cdot)$ denote the largest eigenvalue value function.

2   A combinatorial reformulation of SCCA

This section introduces an equivalent combinatorial optimization reformulation of SCCA. This reformulation serves as the foundation for developing two effective approximation algorithms to solve SCCA in Section 5.

2.1   An exact semidefinite programming representation of CCA

To begin with, let us focus on the classic CCA problem, which refers to SCCA without zero-norm constraints, as defined below:

$$\max_{\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^m} \left\{ \boldsymbol{x}^{\top} \boldsymbol{A} \boldsymbol{y} : \boldsymbol{x}^{\top} \boldsymbol{B} \boldsymbol{x} \leq 1, \boldsymbol{y}^{\top} \boldsymbol{C} \boldsymbol{y} \leq 1 \right\}. \tag{CCA}$$

This formulation of CCA can be regarded as a quadratically constrained quadratic program concerning the variables $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \in \mathbb{R}^{n \times m}$. We next define three-block matrices of size $(n + m)$ below that aid in the presentation of our results.

$$\tilde{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{A}/2 \\ \boldsymbol{A}^{\top}/2 & \boldsymbol{0} \end{pmatrix}, \;\; \tilde{\boldsymbol{B}} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \;\; \tilde{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{pmatrix}.$$

By introducing a size-$(n + m)$ matrix variable $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}^{\top}$ and removing the rank-one constraint on $\boldsymbol{X}$, we can obtain an SDP relaxation of (CCA), as described below

$$\max_{\boldsymbol{X} \in \mathcal{S}_{+}^{m+n}} \left\{ \text{tr}\left(\tilde{\boldsymbol{A}} \boldsymbol{X} \right) : \text{tr}\left(\tilde{\boldsymbol{B}} \boldsymbol{X} \right) \leq 1, \text{tr}\left(\tilde{\boldsymbol{C}} \boldsymbol{X} \right) \leq 1 \right\}. \tag{SDP Relaxation}$$

Next, let us present a key lemma regarding properties of block matrices being positive semidefinite, fundamental for reformulating the SCCA.

Lemma 1 ([12]). *For any symmetric block matrix* $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^{\top} & \boldsymbol{C} \end{pmatrix} \in \mathcal{S}^{n+m}$*, the followings are equivalent:*

(i)   The block matrix is positive semidefinite;
(ii) $\boldsymbol{B} \in \mathcal{S}_{+}^n$, $(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^{\dagger})\boldsymbol{A} = \boldsymbol{0}$, $\boldsymbol{C} - \boldsymbol{A}^{\top}\boldsymbol{B}^{\dagger}\boldsymbol{A} \in \mathcal{S}_{+}^m$; and
(iii) $\boldsymbol{C} \in \mathcal{S}_{+}^m$, $(\boldsymbol{I} - \boldsymbol{C}\boldsymbol{C}^{\dagger})\boldsymbol{A}^{\top} = \boldsymbol{0}$, $\boldsymbol{B} - \boldsymbol{A}\boldsymbol{C}^{\dagger}\boldsymbol{A}^{\top} \in \mathcal{S}_{+}^n$.

Inspired by Lemma 1, we hereby establish the equivalence between CCA and its SDP Relaxation. Remarkably, both of these problems achieve the same optimal value, namely $\sigma_{\max}(\sqrt{\boldsymbol{B}^{\dagger}} \boldsymbol{A} \sqrt{\boldsymbol{C}^{\dagger}})$.

Proposition 1. *For the CCA, we have the following results.*

(i) *Both CCA and its SDP Relaxation have an optimal value $\sigma_{\max}(\sqrt{B^\dagger}A\sqrt{C^\dagger})$;*

(ii) *A pair of optimal solutions (x^*, y^*) to CCA satisfies*

$$x^* = \sqrt{B^\dagger}q, \ \ y^* = \sqrt{C^\dagger}p,$$

where $q \in \mathbb{R}^n, p \in \mathbb{R}^m$ denote a pair of leading singular vectors of matrix $\sqrt{B^\dagger}A\sqrt{C^\dagger}$; and

(iii) *An optimal solution X^* to the SDP Relaxation is*

$$X^* = \begin{pmatrix} x^* \\ y^* \end{pmatrix} \begin{pmatrix} x^* \\ y^* \end{pmatrix}^\top.$$

Proof. See Appendix A.1.   □

The proof of Proposition 1 motivates the following observation on the optimal values of CCA and SCCA.

Observation 1 *The optimal value of CCA is upper bounded by 1, so is the optimal value of SCCA.*

Proof. Since matrix $\begin{pmatrix} B & A \\ A^\top & C \end{pmatrix}$ denotes a covariance matrix of a subset of variables and thus is always positive semidefinite. According to Lemma 1, we have that

$$B \succeq AC^\dagger A^\top \Longrightarrow I \succeq \sqrt{B^\dagger}AC^\dagger A^\top \sqrt{B^\dagger},$$

which means that $\sigma_{\max}\left(\sqrt{B^\dagger}A\sqrt{C^\dagger}\right) \leq 1$ must hold.   □

It is noteworthy that the results presented in Proposition 1 are established through a distinct methodology. This methodology leverages the positive semidefinite condition of block matrices, as shown in Lemma 1, and incorporates duality theory. This approach differs from most prior research [26, 31, 38], which proved Part (i) of Proposition 1 by relying on the singular value decomposition and assuming that matrices B and C are positive definite (i.e., full rank). To the best of our knowledge, [7] showed parts (i) and (ii) of Proposition 1 for a special low-rank CCA problem, where the authors assumed that the covariance matrices are defined as $A = UV^\top$, $B = UU^\top$, and $C = VV^\top$. Remarkably, Proposition 1 extends this result to a more general scenario where B and C are not constrained to be strictly positive definite and A is not constrained to directly depend on B, C, allowing for rank deficiencies and flexible data structure.

2.2   An equivalent formulation of SCCA

In this subsection, we transform SCCA into a combinatorial optimization problem, according to the insights provided by Proposition 1.

Theorem 1. *The SCCA is equivalent to the following combinatorial optimization:*

$$v^* := \max_{\substack{S_1 \subseteq [m], |S_1| \leq s_1, \\ S_2 \subseteq [n], |S_2| \leq s_2}} \left\{ \sigma_{\max}\left(\sqrt{(B_{S_1,S_1})^\dagger}A_{S_1,S_2}\sqrt{(C_{S_2,S_2})^\dagger} \right) \right\}. \qquad (1)$$

Proof. By introducing the subsets (S_1, S_2) to denote the supports of variables $(\boldsymbol{x}, \boldsymbol{y})$ in SCCA, then we can remove the zero-norm constraints on $(\boldsymbol{x}, \boldsymbol{y})$ and reformulate SCCA as

$$v^* := \max_{\substack{S_1 \subseteq [m], |S_1| \leq s_1, \\ S_2 \subseteq [n], |S_2| \leq s_2}} \max_{\substack{\boldsymbol{x} \in \mathbb{R}^{|S_1|}, \\ \boldsymbol{y} \in \mathbb{R}^{|S_2|}}} \left\{ \boldsymbol{x}^\top \boldsymbol{A}_{S_1, S_2} \boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B}_{S_1, S_1} \boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C}_{S_2, S_2} \boldsymbol{y} \leq 1 \right\}.$$

$$(2)$$

Following from the Part (i) in Proposition 1, we can show that for any subsets $S_1 \subseteq [n], S_2 \subseteq [m]$, the following identity holds.

$$\max_{\boldsymbol{x} \in \mathbb{R}^{|S_1|}, \boldsymbol{y} \in \mathbb{R}^{|S_2|}} \left\{ \boldsymbol{x}^\top \boldsymbol{A}_{S_1, S_2} \boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B}_{S_1, S_1} \boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C}_{S_2, S_2} \boldsymbol{y} \leq 1 \right\}$$

$$= \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{S_1, S_1})^\dagger} \boldsymbol{A}_{S_1, S_2} \sqrt{(\boldsymbol{C}_{S_2, S_2})^\dagger} \right).$$

Plugging the result above into the inner maximization problem in (2), we complete the proof.   □

The combinatorial formulation (1) presents significant computational difficulties when attempting to solve SCCA. The primary obstacles are two-fold: first, simultaneously selecting submatrices from the matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ requires a sophisticated optimization across multiple dimensions. Second, the selection criterion is particularly complex, as it involves optimizing the largest singular value of the product of the selected submatrix of $\boldsymbol{A}$ and the square root of pseudo-inverse submatrices of $\boldsymbol{B}$ and $\boldsymbol{C}$. These complexities necessitate effective optimization solution procedures to address the high-dimensional and non-convex nature of the problem.

As a side product of Observation 1, the optimal value of SCCA is trivially upper bounded by 1.

Observation 2 *The optimal value of SCCA satisfies $v^* \leq 1$.*

3 Reformulating SCCA as a mixed-integer semidefinite program (MISDP)

This section formulates an equivalent Mixed-Integer Semidefinite Programming (MISDP) formulation for the SCCA problem. This reformulation serves as the foundation for developing a branch-and-cut algorithm to solve the problem effectively.

3.1 Valid inequalities for SCCA

We prove that there exists a bounded optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ of the SCCA. To be specific, we show that there exists an optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ of the SCCA satisfying the constraints $\|\boldsymbol{x}^*\|_2^2 \leq M_1$ and $\|\boldsymbol{y}^*\|_2^2 \leq M_2$, where M_1 and M_2 are finite-valued parameters.

Proposition 2. *The SCCA admits an optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying $\|\boldsymbol{x}^*\|_2^2 \leq M_1$ and $\|\boldsymbol{y}^*\|_2^2 \leq M_2$, where $M_1 := 1/\lambda_r(\boldsymbol{B}) + 1/(\lambda_r(\boldsymbol{B})s_{\min}(\boldsymbol{B}))$ and $M_2 := 1/\lambda_{\hat{r}}(\boldsymbol{C}) + 1/(\lambda_{\hat{r}}(\boldsymbol{C})s_{\min}(\boldsymbol{C}))$ with $\lambda_r(\boldsymbol{B}), \lambda_{\hat{r}}(\boldsymbol{C})$ being the smallest nonzero eigenvalues of matrices $\boldsymbol{B}, \boldsymbol{C}$ and $s_{\min}(\boldsymbol{R})$ being the smallest nonzero singular value of all the submatrices of the zero eigenvectors of matrix $\boldsymbol{R}$.*

Proof. See Appendix A.2.   □

The proof of Proposition 2 is straightforward in the case when $\boldsymbol{B}$ and $\boldsymbol{C}$ are of full rank as in this case the feasible region is a bounded set. In order to prove the result in the case when $\boldsymbol{B}$ is not full-rank, one has to show that it is possible to construct sparse solutions that are not "too far" away.

In fact, the bounds M_1, M_2 in Proposition 2 also hold for any given feasible subsets (S_1, S_2) of SCCA (1).

Corollary 1. *For any given feasible subsets (S_1, S_2) of SCCA 1, there exists a SCCA feasible solution $(\boldsymbol{x}, \boldsymbol{y})$ such that the supports of $\boldsymbol{x}, \boldsymbol{y}$ are S_1, S_2, respectively and we have that $\|\boldsymbol{x}\|_2^2 \leq M_1$ and $\|\boldsymbol{y}\|_2^2 \leq M_2$, where M_1, M_2 are defined in Proposition 2.*

3.2 An equivalent MISDP formulation

While the combinatorial formulation (1) is elegant in its structure, it poses significant challenges when attempting to solve it to optimality using branch-and-bound based methods. To fill this gap, in this subsection, we derive an equivalent MISDP formulation for SCCA, amenable for developing exact methods.

It is convenient to define the following notation. Let M_{ii} be defined as follows:

$$M_{ii} = \begin{cases} M_1, & \forall i \in [n], \\ M_2, & \forall i \in [n+1, n+m]. \end{cases}$$

Theorem 2. *The SCCA is equivalent to the following MISDP:*

$$v^* := \max_{\substack{\boldsymbol{X} \in \mathcal{S}_+^{n+m}, \\ \boldsymbol{z} \in \mathcal{Z}}} \left\{ \operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{X}) : \operatorname{tr}(\tilde{\boldsymbol{B}}\boldsymbol{X}) \leq 1, \operatorname{tr}(\tilde{\boldsymbol{C}}\boldsymbol{X}) \leq 1, X_{ii} \leq M_{ii} z_i, \forall i \in [n+m] \right\}.$$
(3)

where the feasible set of variables $\boldsymbol{z}$ is defined as $\mathcal{Z} := \{\boldsymbol{z} \in \{0,1\}^{n+m} : \sum_{i \in [n]} z_i \leq s_1, \sum_{i \in [n+1, n+m]} z_i \leq s_2\}$.

Proof. For the SCCA (2), according to Proposition 1, the inner maximization problem admits an exact semidefinite programming formulation. Using the variables $\boldsymbol{z} \in \mathcal{Z}$ to describe the set constraints in SCCA (2), we can reformulate it as

$$v^* := \max_{\boldsymbol{z} \in \mathcal{Z}} \max_{\boldsymbol{X} \in \mathcal{S}_+^{n+m}} \big\{ \operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{X}) : \operatorname{tr}(\tilde{\boldsymbol{B}}\boldsymbol{X}) \leq 1, \operatorname{tr}(\tilde{\boldsymbol{C}}\boldsymbol{X}) \leq 1,$$

$$X_{ii}(1 - z_i) = 0, \forall i \in [m+n] \big\}.$$
(4)

Proposition 2 shows that there is an optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ to SCCA that satisfies $\|\boldsymbol{x}^*\|_2^2 \le M_1$ and $\|\boldsymbol{y}^*\|_2^2 \le M_2$. Based on this, we can construct an optimal solution $(\boldsymbol{z}^*, \boldsymbol{X}^*)$ for SCCA (4) by letting

$$\boldsymbol{X}^* = \begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}^\top, z_i = \begin{cases} 1 & \text{if } x_i^* \ne 0 \\ 0 & \text{if } x_i^* = 0 \end{cases}, \forall i \in [n], z_{i+n} = \begin{cases} 1 & \text{if } y_i^* \ne 0 \\ 0 & \text{if } y_i^* = 0 \end{cases}, \forall i \in [m],$$

where the optimal solution $\boldsymbol{X}^*$ satisfies the following inequalities

$$X_{ii}^* = (x_i^*)^2 \le M_1 z_i, \forall i \in [n], \ \ X_{(i+n)(i+n)}^* = (y_i^*)^2 \le M_2 z_{i+n}, \forall i \in [m].$$

This allows us to recast the SCCA (4) into an MISDP formulation (3). $\qquad\square$

Note that the proposed MISDP formulation (3) is of size $(n+m) \times (n+m)$ since our matrix variable $\boldsymbol{X}$ replaces $\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}^\top$ in SCCA. Relaxing the binary variables in SCCA (3) to be continuous, we obtain an upper bound of SCCA (3), i.e., $v^* \le \widehat{v}$

$$\widehat{v} := \max_{\boldsymbol{X} \in \mathcal{S}_+^{n+m}, \boldsymbol{z} \in \widehat{\mathcal{Z}}} \{ \operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{X}) : \operatorname{tr}(\tilde{\boldsymbol{B}}\boldsymbol{X}) \le 1, \operatorname{tr}(\tilde{\boldsymbol{C}}\boldsymbol{X}) \le 1, X_{ii} \le M_{ii} z_i, \forall i \in [n+m] \}.$$
(5)

where $\widehat{\mathcal{Z}} := \{ \boldsymbol{z} \in [0,1]^{n+m} : \sum_{i \in [n]} z_i \le s_1, \sum_{i \in [n+1, n+m]} z_i \le s_2 \}$. This SDP relaxation (5) can be directly solved by commercial solvers such as MOSEK or SDPT3.

3.3   Developing a branch-and-cut algorithm with closed-form cuts

By dualizing the inner maximization problem over $\boldsymbol{X}$ in the MISDP (3), in this subsection, we derive an equivalent mixed-integer linear program for SCCA, which motivates us to develop a branch-and-cut algorithm.

By separating the binary variables $\boldsymbol{z}$, we rewrite the MISDP (3) as

$$v^* := \max_{\boldsymbol{z} \in \mathcal{Z}, v} \{ v : v \le f(\boldsymbol{z}) \},$$
(6)

where the function $f(\boldsymbol{z})$ is defined as

$$f(\boldsymbol{z}) := \max_{\boldsymbol{X} \in \mathcal{S}_+^{n+m}} \left\{ \operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{X}) : \operatorname{tr}(\tilde{\boldsymbol{B}}\boldsymbol{X}) \le 1, \operatorname{tr}(\tilde{\boldsymbol{C}}\boldsymbol{X}) \le 1, X_{ii} \le M_{ii} z_i, \forall i \in [n+m] \right\}.$$
(7)

By introducing the Lagrangian multipliers $(\theta_1, \theta_2, \boldsymbol{\lambda})$, the Lagrangian dual of the maximization problem (7) can be written as

$$f(\boldsymbol{z}) = \min_{\substack{\theta_1 \ge 0, \theta_2 \ge 0, \\ \boldsymbol{\lambda} \in \mathbb{R}_+^{n+m}}} \max_{\boldsymbol{X} \in \mathcal{S}_+^{n+m}} \operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{X}) - \theta_1 \operatorname{tr}(\tilde{\boldsymbol{B}}\boldsymbol{X}) - \theta_2 \operatorname{tr}(\tilde{\boldsymbol{C}}\boldsymbol{X}) + \theta_1 + \theta_2,$$

$$- \sum_{i \in [n+m]} \lambda_i X_{ii} + \sum_{i \in [n+m]} \lambda_i M_{ii} z_i$$

$$= \min_{\substack{\theta_1 \ge 0, \theta_2 \ge 0, \\ \boldsymbol{\lambda} \in \mathbb{R}_+^{n+m}}} \left\{ \theta_1 + \theta_2 + \sum_{i \in [n+m]} \lambda_i M_{ii} z_i : \begin{pmatrix} \theta_1 \boldsymbol{B} & -\boldsymbol{A}/2 \\ -\boldsymbol{A}^\top/2 & \theta_2 \boldsymbol{C} \end{pmatrix} \succeq -\operatorname{Diag}(\boldsymbol{\lambda}) \right\},$$
(8)

where the strong duality holds due to the function $f(\boldsymbol{z})$ being concave, bounded, and thus continuous in the set $\widehat{\mathcal{Z}}$ and Slater condition holds for any interior point $\boldsymbol{z}$ in the set $\widehat{\mathcal{Z}}$.

Below, we derive the closed-form expression of the function $f(\boldsymbol{z})$ with the given binary variable $\boldsymbol{z} \in \mathcal{Z}$. This allows us to reformulate SCCA (6) as a mixed-integer linear program with exponentially many linear constraints and an efficient separation oracle.

Proposition 3. *The SCCA (6) is equivalent to*

$$
v^* := \max_{\boldsymbol{z} \in \mathcal{Z}, v} \left\{ v : v \leq \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right) + \right.
$$
$$
\left. \sum_{i \in S_1 \cup (S_2 + n)} \lambda^* M_{ii} z_i : \forall S_1 \subseteq [n], |S_1| \leq s_1, S_2 \subseteq [m], |S_2| \leq s_2 \right\}, \tag{9}
$$

where for a pair of subsets (S_1, S_2), the scalar λ^* is defined as the largest positive eigenvalue of matrix $\boldsymbol{D}_2^\top \boldsymbol{D}_1^{-1} \boldsymbol{D}_2 - \boldsymbol{D}_3$ with

$$
\boldsymbol{D}_1 := \begin{pmatrix} \theta_1^* \boldsymbol{B}_{S_1,S_1} & -\boldsymbol{A}_{S_1,S_2}/2 \\ -\boldsymbol{A}_{S_1,S_2}^\top/2 & \theta_2^* \boldsymbol{C}_{S_2,S_2} \end{pmatrix}, \quad \boldsymbol{D}_2 := \begin{pmatrix} \theta_1^* \boldsymbol{B}_{S_1,[n]\setminus S_1} & -\boldsymbol{A}_{S_1,[m]\setminus S_2}/2 \\ -\boldsymbol{A}_{S_2,[n]\setminus S_1}^\top/2 & \theta_2^* \boldsymbol{C}_{S_2,[m]\setminus S_2} \end{pmatrix},
$$

and

$$
\boldsymbol{D}_3 := \begin{pmatrix} \theta_1^* \boldsymbol{B}_{[n]\setminus S_1,[n]\setminus S_1} & -\boldsymbol{A}_{[n]\setminus S_1,[m]\setminus S_2}/2 \\ -\boldsymbol{A}_{[n]\setminus S_1,[m]\setminus S_2}^\top/2 & \theta_2^* \boldsymbol{C}_{[m]\setminus S_2,[m]\setminus S_2} \end{pmatrix},
$$

where $\theta_1^* = \theta_2^* = \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right) /2$.

Proof. See Appendix A.3. □

We note that SCCA (9) can be implemented via a delayed cut-generation procedure. That is, at each feasible branch-and-bound node with a binary solution $\widehat{\boldsymbol{z}}$, let $S_1 := \{i : \widehat{z}_i = 1, \forall i \in [n]\}$ and $S_2 := \{i - n : \widehat{z}_i = 1, \forall i \in [n+1, n+m]\}$. Then we can compute the corresponding scalar λ^* and generate the following valid inequality based on (9):

$$
v \leq \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right) + \sum_{i \in S_1 \cup (S_2 + n)} \lambda^* M_{ii} z_i.
$$

4 Low-rank SCCA

In practice, it is common that the sample covariance matrix $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{C} \end{pmatrix}$ exhibits low-rank characteristics. This phenomenon is especially prominent when dealing with high-dimensional, low-sample size data (e.g., gene expression data [35]). In this section, we study two special cases of low-rank SCCA and their computational complexities.

4.1   Special Case I: SCCA with low-rank covariance matrices

In this section, we show that the computational complexity of SCCA is contingent upon the ranks of the covariance matrices $\boldsymbol{B}$ and $\boldsymbol{C}$. To be more precise, when the sparsity level s_1 (or s_2) is equal to or greater than the rank of the covariance matrix $\boldsymbol{B}$ (or $\boldsymbol{C}$), the imposition of a zero-norm constraint over $\boldsymbol{x}$ (or $\boldsymbol{y}$) in SCCA becomes redundant. Consequently, lower ranks in the covariance matrices correspond to better computational complexity in solving SCCA.

Theorem 3. *Suppose $r := \mathrm{rank}(\boldsymbol{B})$ and $\widehat{r} := \mathrm{rank}(\boldsymbol{C})$, then the SCCA takes a complexity of $\mathcal{O}(n^{r-1}m^{\widehat{r}-1} + n^{r-1} + m^{\widehat{r}-1})$. The following results hold:*

(i) When $s_1 \geq r$ and $s_2 \geq \widehat{r}$, the SCCA problem is equivalent to CCA, i.e.,

$$v^* := \max_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^m} \left\{ \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y} \leq 1 \right\}; \tag{10}$$

(ii) When $s_1 \geq r$ and $s_2 < \widehat{r}$, the SCCA problem can be reduced to

$$v^* := \max_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^m} \left\{ \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y} \leq 1, \|\boldsymbol{y}\|_0 \leq s_2 \right\}; \tag{11}$$

(iii) When $s_1 < r$ and $s_2 \geq \widehat{r}$, the SCCA problem can be reduced to

$$v^* := \max_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^m} \left\{ \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y} \leq 1, \|\boldsymbol{x}\|_0 \leq s_1 \right\}. \tag{12}$$

Proof. See Appendix A.4. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

The proof of Theorem 3 implies that CCA admits an optimal sparse solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying $\|\boldsymbol{x}^*\|_0 \leq r$ and $\|\boldsymbol{y}^*\|_0 \leq \widehat{r}$, provided that $\boldsymbol{B}, \boldsymbol{C}$ are of rank-r, $\widehat{r}$, respectively. Thus, Theorem 3 establishes a sufficient condition (i.e., $s_1 \leq r, s_2 \leq \widehat{r}$) about when CCA can be equivalent to SCCA. Besides, Theorem 3 implies the complexity of solving SCCA, as summarized below.

Corollary 2. *Suppose $r := \mathrm{rank}(\boldsymbol{B})$ and $\widehat{r} := \mathrm{rank}(\boldsymbol{C})$. There exists an algorithm that can find an optimal solution to SCCA in $\mathcal{O}(n^{r-1}m^{\widehat{r}-1})$ time complexity.*

4.2   Special Case II: SCCA with a rank-one cross-covariance matrix

In this subsection, we study the other interesting low-rank special case of SCCA where the cross-covariance matrix $\boldsymbol{A}$ is rank-one. For this special case, we prove its NP-hardness with reduction to the sparse regression problem.

We observe that SCCA can be separable over variables $\boldsymbol{x}$ and $\boldsymbol{y}$ for the rank-one $\boldsymbol{A}$. In fact, suppose that $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{b}^\top$, then SCCA is equivalent to

$$v^* := \max_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^m} \left\{ \boldsymbol{x}^\top \boldsymbol{a}\boldsymbol{b}^\top \boldsymbol{y} : \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} \leq 1, \boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y} \leq 1, \|\boldsymbol{x}\|_0 \leq s_1, \|\boldsymbol{y}\|_0 \leq s_2 \right\} \tag{13}$$

which can be equivalently the product of the optimal values of the following two subproblems:

$$\begin{aligned} v_x &:= \max_{\boldsymbol{x}\in\mathbb{R}^n} \{ \boldsymbol{a}^\top \boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} \leq 1, \|\boldsymbol{x}\|_0 \leq s_1 \}, \\ v_y &:= \max_{\boldsymbol{y}\in\mathbb{R}^m} \{ \boldsymbol{b}^\top \boldsymbol{y} : \boldsymbol{y}^\top \boldsymbol{C}\boldsymbol{y} \leq 1, \|\boldsymbol{y}\|_0 \leq s_2 \}. \end{aligned} \tag{14}$$

That is, the identity $v^* = v_x v_y$ holds. According to Proposition 2, introducing binary variables, we can reformulate two subproblems (5) as mixed-integer convex quadratic programs. Consequently, the rank-one SCCA problem, as formulated in (13), simplifies to two mixed-integer convex quadratic programs. This simplification is much more tractable compared to addressing the MISDP (3), which involves a large-sized positive semidefinite variable $\boldsymbol{X}$ of dimension $(n+m) \times (n+m)$. Our numerical findings confirm the reduced complexity of the rank-one SCCA model.

Next, we show that each subproblem in (14) can be reduced to the classic sparse regression problem [1, 27] and is thus NP-hard as shown below.

Theorem 4. *When matrix $\boldsymbol{A} := \boldsymbol{a}^\top \boldsymbol{b}$ is rank-one, each maximization problem in* (14) *is NP-hard.*

Proof. See Appendix A.5. □

Theorem 4 links the maximization problem (14) and the well-known sparse regression problem, implying that even solving the rank-one SCCA problem (13) is NP-hard. This also suggests employing strong perspective formulations (see, e.g., [1, 37]) when solving the subproblems (14), which are shown to be stronger and easier to solve than the SDP relaxation (5) in our numerical study.

5 Numerical results

This section tests the numerical performance of our formulations and algorithms on synthetic data. All the experiments are conducted in Python 3.6 with calls to Gurobi 9.5.2 and MOSEK 10.0.29 on a PC with 10-core CPU, 16-core GPU, and 16GB of memory.

We generate random instances by fixing the dimensions n, m and the sparsity levels s_1, s_2. For each instance, given parameters (n, m, s_1, s_2), we first generate the covariance matrix $\begin{pmatrix} \boldsymbol{B}^0 & \boldsymbol{A}^0 \\ (\boldsymbol{A}^0)^\top & \boldsymbol{C}^0 \end{pmatrix}$ as follows;

(i) $\boldsymbol{B}^0 \in \mathcal{S}_{++}^n$: Let $\widehat{\boldsymbol{B}}$ consist of $n \times n$ elements generated from a normal distribution $\mathcal{N}(0, 1)$, and let $\boldsymbol{B}^0 = \widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top + \boldsymbol{I}$;

(ii) $\boldsymbol{C}^0 \in \mathcal{S}_{++}^m$: Let $\widehat{\boldsymbol{C}}$ consist of $m \times m$ elements generated from a normal distribution $\mathcal{N}(0, 1)$, and let $\boldsymbol{C}^0 = \widehat{\boldsymbol{C}}\widehat{\boldsymbol{C}}^\top + \boldsymbol{I}$; and

(iii) $\boldsymbol{A}^0 \in \mathbb{R}^{n \times m} := \lambda \boldsymbol{B}^0 \boldsymbol{u}\boldsymbol{v}^\top \boldsymbol{C}^0$: We generate λ uniformly from $(0, 1)$, and vectors $\boldsymbol{u}, \boldsymbol{v}$ are generated from a normal distribution $\mathcal{N}(0, 1)$ that satisfy $\|\boldsymbol{u}\|_0 = s_1, \|\boldsymbol{v}\|_0 = s_2, \boldsymbol{u}^\top \boldsymbol{B}^0 \boldsymbol{u} = 1$ and $\boldsymbol{v}^\top \boldsymbol{C}^0 \boldsymbol{v} = 1$.

Next, we sample $N = 5,000$ data points $\{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i \in [N]} \in \mathbb{R}^n \times \mathbb{R}^m$ from the normal distribution with zero mean and the covariance $\begin{pmatrix} \boldsymbol{B}^0 & \boldsymbol{A}^0 \\ (\boldsymbol{A}^0)^\top & \boldsymbol{C}^0 \end{pmatrix}$. Then, let us estimate $\boldsymbol{A}^0, \boldsymbol{B}^0, \boldsymbol{C}^0$ by sample covariance matrices below

$$\boldsymbol{A} = \sum_{i \in [N]} \boldsymbol{u}_i \boldsymbol{v}_i^\top, \quad \boldsymbol{B} = \sum_{i \in [N]} \boldsymbol{u}_i \boldsymbol{u}_i^\top, \quad \boldsymbol{C} = \sum_{i \in [N]} \boldsymbol{v}_i \boldsymbol{v}_i^\top.$$

The numerical results are presented in Table 1 that include multiple instances with various parameters (n, m, s_1, s_2). Throughout, the computational time is in seconds, the time limit is one hour, and the dashed line "-" denotes the unsolved case within the time limit. First, based on the combinatorial formulation (1), we consider using the greedy and local search algorithms to approximately solve SCCA, and their detailed implementation can be found in Appendix B. Note that we let **LB** denote the lower bound obtained from the approximation algorithm. In Table 1, we define **gap(%)**:= $(\widehat{v} - v^*)/v^*$ to be the optimality gap of the upper bound in (5), and we replace v^* with the best lower bound when v^* is not available. It is seen that the greedy and local search algorithms are quite scalable, and the SDP relaxation (5) yields a tight upper bound with an optimality gap at most 8.16%. We apply a branch-and-cut algorithm to solve SCCA (9) via the delayed cut generation procedure, which can handle the case up to size 20 in Table 1. One reason may be because SCCA (9) has a weak relaxation bound. Therefore, although our proposed cut in Section 3 admits closed form, the branch-and-cut algorithm explores a considerable amount of nodes before termination.

Table 1. Solving SCCA with synthetic data

n	m	s_1	s_2	Greedy		Local search		SDP relaxation (5)			SCCA (9)	
				LB	time(s)	LB	time(s)	$\widehat{v}$	gap(%)	time(s)	v^*	time(s)
10	10	5	5	0.244	1	0.244	1	0.247	1.33	1	0.244	26
20	20	5	5	0.244	1	0.244	1	0.256	1.23	1	0.244	2217
20	20	10	10	0.275	1	0.275	1	0.278	1.23	1	0.275	3562
40	40	5	5	0.695	1	0.695	1	0.701	0.83	1	-	-
40	40	10	10	0.705	1	0.705	1	0.708	0.45	1	-	-
40	60	5	10	0.707	1	0.707	1	0.714	0.93	1	-	-
40	60	10	5	0.704	1	0.704	1	0.708	0.65	1	-	-
60	60	5	5	0.720	1	0.720	1	0.727	0.86	14	-	-
60	60	10	10	0.714	1	0.714	1	0.721	1.00	12	-	-
80	80	5	5	0.395	1	0.395	1	0.427	8.16	56	-	-
80	80	10	10	0.399	1	0.399	1	0.428	7.36	62	-	-
100	100	5	5	0.942	1	0.942	1	0.944	0.23	257	-	-
100	100	10	10	0.940	1	0.940	1	0.942	0.23	313	-	-
120	120	5	5	0.479	1	0.479	1	0.517	7.90	1360	-	-
120	120	10	10	0.501	1	0.501	1	0.942	7.86	1569	-	-

The complexity analysis of low-rank SCCA in Section 4 indicates that rank-one SCCA (13) can be more tractable, as we decompose it into two subproblems in (14). By approximating $\boldsymbol{A}$ with a rank-one matrix that consists of leading singular value and vectors, Table 2 presents the numerical results for solving rank-one SCCA (13). In addition to the SDP relaxation (5), we consider the strong perspective formulations of subproblems (14) to provide an upper bound for rank-one SCCA (13) (see, e.g., [1, 37]), denoted by **Perspective** in Table 2. We also compute its optimality gap and compare it with SDP relaxation (5). It is obvious that perspective relaxation is computationally efficient and yields smaller optimality gaps, which solves all the testing cases in 15 seconds with an optimality gap of up to 11.3%. As previously mentioned in Section 4.2, we can solve two mixed-integer quadratic programs below via Gurobi to find the optimal value of rank-one SCCA

(13), i.e., $v^* := v_x v_y$, where the performance can be found in the last column of Section 4.2. We see that we can solve size-100×100 rank-one SCCA (13).

Table 2. Solving rank-one SCCA with synthetic data

n	m	s_1	s_2	Greedy		Local Search		SDP relaxtion (5)		Perspective		SCCA (13)	
				LB	time(s)	LB	time(s)	gap(%)	time(s)	gap(%)	time(s)	v^*	time(s)
50	50	10	10	0.382	1	0.382	1	3.79	6	2.44	1	0.382	30
50	50	20	20	0.409	1	0.409	1	2.81	7	1.74	1	0.409	293
100	100	10	10	0.928	1	0.928	1	0.79	492	0.47	1	0.928	81
100	100	20	20	0.943	1	0.943	2	0.49	685	0.31	1	0.943	3463
200	200	10	10	0.549	1	0.549	1	-	-	7.38	1	-	-
200	200	20	20	0.524	1	0.524	5	-	-	9.70	1	-	-
300	300	10	10	0.874	1	0.874	1	-	-	2.56	6	-	-
300	300	20	20	0.878	1	0.878	9	-	-	2.49	8	-	-
400	400	10	10	0.840	1	0.840	2	-	-	4.43	9	-	-
400	400	20	20	0.842	1	0.842	14	-	-	4.34	10	-	-
500	500	10	10	0.701	1	0.701	2	-	-	11.3	14	-	-
500	500	20	20	0.710	6	0.710	59	-	-	10.9	15	-	-

References

[1] Atamturk, A., Gomez, A.: Rank-one convexification for sparse regression. arXiv preprint arXiv:1901.10334 (2019)

[2] Bertsimas, D., Cory-Wright, R.: Solving large-scale sparse PCA to certifiable (near) optimality. The Journal of Machine Learning Research 23(1), 566–600 (2022)

[3] Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. The Annals of Statistics 44(2), 813 – 852 (2016), https://doi.org/10.1214/15-AOS1388

[4] Chan, S.O., Papailliopoulos, D., Rubinstein, A.: On the approximability of sparse PCA. In: Conference on Learning Theory. pp. 623–646. PMLR (2016)

[5] Chen, M., Gao, C., Ren, Z., Zhou, H.H.: Sparse cca via precision adjusted iterative thresholding. arXiv preprint arXiv:1311.6186 (2013)

[6] Chowdhury, A., Drineas, P., Woodruff, D.P., Zhou, S.: Approximation algorithms for sparse principal component analysis. arXiv preprint arXiv:2006.12748 (2020)

[7] Chu, D., Liao, L.Z., Ng, M.K., Zhang, X.: Sparse canonical correlation analysis: New formulation and algorithm. IEEE transactions on pattern analysis and machine intelligence 35(12), 3050–3065 (2013)

[8] d'Aspremont, A., Bach, F., El Ghaoui, L.: Optimal solutions for sparse principal component analysis. Journal of Machine Learning Research 9(7) (2008)

[9] d'Aspremont, A., Ghaoui, L., Jordan, M., Lanckriet, G.: A direct formulation for sparse PCA using semidefinite programming. Advances in neural information processing systems 17 (2004)

[10] Dey, S.S., Mazumder, R., Wang, G.: Using ℓ_1-relaxation and integer programming to obtain dual bounds for sparse PCA. Operations Research 70(3), 1914–1932 (2022)

[11] Dey, S.S., Molinaro, M., Wang, G.: Solving sparse principal component analysis with global support. Mathematical Programming 199(1-2), 421–459 (2023)

[12] Gallier, J., et al.: The schur complement and symmetric positive semidefinite (and definite) matrices (2019). URL https://www. cis. upenn. edu/jean/schur-comp. pdf (2020)

[13] Gao, C., Ma, Z., Ren, Z., Zhou, H.H.: Minimax estimation in sparse canonical correlation analysis (2015)

[14] Gao, C., Ma, Z., Zhou, H.H.: Sparse cca: Adaptive estimation and computational barriers (2017)

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

[16] Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE transactions on speech and audio processing 2(4), 578–589 (1994)

[17] Hocking, R.R., Leslie, R.: Selection of the best subset in regression analysis. Technometrics 9(4), 531–540 (1967)

[18] Hotelling, H.: The most predictable criterion. Journal of educational Psychology 26(2), 139 (1935)

[19] Huang, H., He, H., Fan, X., Zhang, J.: Super-resolution of human face image using canonical correlation analysis. Pattern Recognition 43(7), 2532–2543 (2010)

[20] Jeffers, J.N.: Two case studies in the application of principal component analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics) 16(3), 225–236 (1967)

[21] Lê Cao, K.A., Martin, P.G., Robert-Granié, C., Besse, P.: Sparse canonical methods for biological data integration: application to a cross-platform study. BMC bioinformatics 10, 1–17 (2009)

[22] Li, Y., Xie, W.: Exact and approximation algorithms for sparse PCA. arXiv preprint arXiv:2008.12438 (2020)

[23] Li, Y., Xie, W.: Beyond symmetry: Best submatrix selection for the sparse truncated svd. arXiv preprint arXiv:2105.03179 (2021)

[24] Lin, D., Calhoun, V.D., Wang, Y.P.: Correspondence between fmri and snp data by group sparse canonical correlation analysis. Medical image analysis 18(6), 891–902 (2014)

[25] Lin, D., Zhang, J., Li, J., Calhoun, V.D., Deng, H.W., Wang, Y.P.: Group sparse canonical correlation analysis for genomic data integration. BMC bioinformatics 14(1), 1–16 (2013)

[26] Lu, Y., Foster, D.P.: Large scale canonical correlation analysis with iterative least squares. Advances in Neural Information Processing Systems 27 (2014)

[27] Miller, A.: Subset selection in regression. CRC Press (2002)

[28] Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM journal on computing 24(2), 227–234 (1995)

[29] Papailiopoulos, D., Dimakis, A., Korokythakis, S.: Sparse PCA through low-rank approximations. In: International Conference on Machine Learning. pp. 747–755. PMLR (2013)

[30] Parkhomenko, E., Tritchler, D., Beyene, J.: Genome-wide sparse canonical correlation of gene expression with genotypes. In: BMC proceedings. vol. 1, pp. 1–5. BioMed Central (2007)

[31] Parkhomenko, E., Tritchler, D., Beyene, J.: Sparse canonical correlation analysis with application to genomic data integration. Statistical applications in genetics and molecular biology 8(1) (2009)

[32] Vinokourov, A., Cristianini, N., Shawe-Taylor, J.: Inferring a semantic representation of text via cross-language correlation analysis. Advances in neural information processing systems 15 (2002)

[33] Waaijenborg, S., Verselewel de Witt Hamer, P.C., Zwinderman, A.H.: Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis. Statistical applications in genetics and molecular biology 7(1) (2008)

[34] Watanabe, A., Tamura, R., Takano, Y., Miyashiro, R.: Branch-and-bound algorithm for optimal sparse canonical correlation analysis. Expert Systems with Applications 217, 119530 (2023)

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

[36] Witten, D.M., Tibshirani, R.J.: Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology 8(1) (2009)

[37] Xie, W., Deng, X.: Scalable algorithms for the sparse ridge regression. SIAM Journal on Optimization 30(4), 3359–3386 (2020)

[38] Yang, X., Liu, W., Liu, W., Tao, D.: A survey on canonical correlation analysis. IEEE Transactions on Knowledge and Data Engineering 33(6), 2349–2368 (2019)

Appendix A: Proofs

A.1 Proof of Proposition 1

Proof. The proof includes three parts.

Part (i). To prove the equivalence between CCA and its SDP Relaxation, let us introduce the Lagrangian multiplies $\theta_1 \geq 0, \theta_2 \geq 0$ corresponding to two constraints in SDP Relaxation, which leads to the following Lagrangian dual problem

$$\min_{\theta_1 \geq 0, \theta_2 \geq 0} \left\{ \theta_1 + \theta_2 : \theta_1 \tilde{\boldsymbol{B}} + \theta_2 \tilde{\boldsymbol{C}} \succeq \tilde{\boldsymbol{A}} \right\} = \min_{\theta_1 \geq 0, \theta_2 \geq 0} \left\{ \theta_1 + \theta_2 : \begin{pmatrix} \theta_1 \boldsymbol{B} & \frac{\boldsymbol{A}}{-2} \\ \frac{\boldsymbol{A}^\top}{-2} & \theta_2 \boldsymbol{C} \end{pmatrix} \succeq 0 \right\}$$
$$(15)$$

where the equation results from the definition of block matrices $\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{B}}$, and $\tilde{\boldsymbol{C}}$. Given the nonzero matrices $\boldsymbol{A} \neq \boldsymbol{0}, \boldsymbol{B} \neq \boldsymbol{0}, \boldsymbol{C} \neq \boldsymbol{0}$ and positive semidefinite matrices $\boldsymbol{B} \succeq 0, \boldsymbol{C} \succeq 0$, following Lemma 1, we must have $\theta_2 \boldsymbol{C} - \boldsymbol{A}^\top (\theta_1 \boldsymbol{B})^\dagger \boldsymbol{A}/4 \succeq 0$ and $\theta_1 \boldsymbol{B} - \boldsymbol{A}(\theta_2 \boldsymbol{C})^\dagger \boldsymbol{A}^\top/4 \succeq 0$, implying that either $\theta_1 = 0$ or $\theta_2 = 0$ is infeasible to the minimization problem above. That is, $\theta_1 > 0$ and $\theta_2 > 0$ must hold.

According to Lemma 1, the block matrix $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{C} \end{pmatrix}$ is positive semidefinite, implying that $(\boldsymbol{I} - \boldsymbol{C}\boldsymbol{C}^\dagger)\boldsymbol{A}^\top = \boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^\dagger)\boldsymbol{A} = \boldsymbol{0}$. Then, it is easy to show

$$\left(\boldsymbol{I} - \theta_2 \boldsymbol{C}(\theta_2 \boldsymbol{C})^\dagger \right) \frac{\boldsymbol{A}^\top}{2} = \boldsymbol{0}, \forall \theta_2 > 0.$$

Given $\theta_1, \theta_2 > 0$ and using Lemma 1, the result above allows us to further simplify the right-hand side minimization problem in (15) to

$$\min_{\theta_1 \geq 0, \theta_2 \geq 0} \left\{ \theta_1 + \theta_2 : 4\theta_1 \theta_2 \boldsymbol{B} \succeq \boldsymbol{A}\boldsymbol{C}^\dagger \boldsymbol{A}^\top \right\}$$
$$= \min_{\theta_1 \geq 0, \theta_2 \geq 0} \left\{ \theta_1 + \theta_2 : 4\theta_1 \theta_2 \geq \sigma_{\max}^2 \left(\sqrt{\boldsymbol{B}^\dagger} \boldsymbol{A} \sqrt{\boldsymbol{C}^\dagger} \right) \right\} = \sigma_{\max} \left(\sqrt{\boldsymbol{B}^\dagger} \boldsymbol{A} \sqrt{\boldsymbol{C}^\dagger} \right),$$

where the first equation is because

$$4\theta_1 \theta_2 \boldsymbol{B} \succeq \boldsymbol{A}\boldsymbol{C}^\dagger \boldsymbol{A}^\top \iff 4\theta_1 \theta_2 \boldsymbol{I} \succeq \sqrt{\boldsymbol{\Lambda}^{-1}} \boldsymbol{Q}^\top \boldsymbol{A}\boldsymbol{C}^\dagger \boldsymbol{A}^\top \boldsymbol{Q} \sqrt{\boldsymbol{\Lambda}^{-1}}$$
$$\iff 4\theta_1 \theta_2 \geq \lambda_{\max} \left(\sqrt{\boldsymbol{\Lambda}^{-1}} \boldsymbol{Q}^\top \boldsymbol{A}\boldsymbol{C}^\dagger \boldsymbol{A}^\top \boldsymbol{Q} \sqrt{\boldsymbol{\Lambda}^{-1}} \right)$$
$$\iff 4\theta_1 \theta_2 \geq \lambda_{\max} \left(\sqrt{\boldsymbol{C}^\dagger} \boldsymbol{A}^\top \boldsymbol{B}^\dagger \boldsymbol{A} \sqrt{\boldsymbol{C}^\dagger} \right) \iff 4\theta_1 \theta_2 \geq \sigma_{\max}^2 \left(\sqrt{\boldsymbol{B}^\dagger} \boldsymbol{A} \sqrt{\boldsymbol{C}^\dagger} \right),$$

where we let $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ denote the eigendecomposition of matrix $\boldsymbol{B}$ with $\boldsymbol{\Lambda}$ containing all the positive eigenvalues.

As a result, the dual problem of SDP Relaxation admits an optimal value of $\sigma_{\max} \left(\sqrt{\boldsymbol{B}^\dagger} \boldsymbol{A} \sqrt{\boldsymbol{C}^\dagger} \right)$, which gives an upper bound of the CCA and its SDP Relaxation. Next, we construct their optimal solutions, which exactly attain this upper bound. Thus, this upper bound is achievable and equals their optimal values.

Part (ii). For the CCA, let us consider a part of optimal solutions $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ below

$$\boldsymbol{x}^* = \sqrt{\boldsymbol{B}^\dagger}\boldsymbol{q}, \ \ \boldsymbol{y}^* = \sqrt{\boldsymbol{C}^\dagger}\boldsymbol{p},$$

with $\boldsymbol{q} \in \mathbb{R}^n, \boldsymbol{p} \in \mathbb{R}^m$ denoting a pair of leading singular vectors of matrix $\sqrt{\boldsymbol{B}^\dagger}\boldsymbol{A}\sqrt{\boldsymbol{C}^\dagger}$.

First, $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is feasible to the CCA as

$$(\boldsymbol{x}^*)^\top \boldsymbol{B}\boldsymbol{x}^* = \boldsymbol{q}^\top \sqrt{\boldsymbol{B}^\dagger}\boldsymbol{B}\sqrt{\boldsymbol{B}^\dagger}\boldsymbol{q} \leq \boldsymbol{q}^\top \boldsymbol{q} = 1, \ (\boldsymbol{y}^*)^\top \boldsymbol{C}\boldsymbol{y}^* = \boldsymbol{p}^\top \sqrt{\boldsymbol{C}^\dagger}\boldsymbol{C}\sqrt{\boldsymbol{C}^\dagger}\boldsymbol{p} \leq \boldsymbol{p}^\top \boldsymbol{p} = 1,$$

where the inequalities stem from the facts that $\boldsymbol{I} \succeq \sqrt{\boldsymbol{B}^\dagger}\boldsymbol{B}\sqrt{\boldsymbol{B}^\dagger}$ and $\boldsymbol{I} \succeq \sqrt{\boldsymbol{C}^\dagger}\boldsymbol{C}\sqrt{\boldsymbol{C}^\dagger}$.

On the other hand, according to the definitions of $\boldsymbol{q}, \boldsymbol{p}$, we can show that $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is optimal to the CCA, i.e.,

$$(\boldsymbol{x}^*)^\top \boldsymbol{A}\boldsymbol{y}^* = \boldsymbol{q}^\top \sqrt{\boldsymbol{B}^\dagger}\boldsymbol{A}\sqrt{\boldsymbol{C}^\dagger}\boldsymbol{p} = \sigma_{\max}\left(\sqrt{\boldsymbol{B}^\dagger}\boldsymbol{A}\sqrt{\boldsymbol{C}^\dagger}\right).$$

Part (iii). In a similar vein, we can show that $\boldsymbol{X}^* = \begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}\begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}^\top$ is optimal to SDP Relaxation with the optimal value $\sigma_{\max}\left(\sqrt{\boldsymbol{B}^\dagger}\boldsymbol{A}\sqrt{\boldsymbol{C}^\dagger}\right)$. $\qquad\square$

A.2 Proof of Proposition 2

Proof. Let $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ denote an optimal solution to SCCA. We bound $\|\boldsymbol{x}^*\|_2$ first and the same technique can be also straightforwardly applied to bound $\|\boldsymbol{y}^*\|_2$.

For matrix $\boldsymbol{B} \in \mathcal{S}_+^n$ of rank r, we let $\{\boldsymbol{q}_i\}_{i \in [n]} \in \mathbb{R}^n$ denote the eigenvectors corresponding to n eigenvalues $\boldsymbol{\lambda}$ of $\boldsymbol{B}$ such that $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0$. Thus, $\{\boldsymbol{q}_i\}_{i \in [n]}$ are orthonormal and span the space of $\mathbb{R}^n$. Hence, there exists $\boldsymbol{\alpha} \in \mathbb{R}^n$ such that $\boldsymbol{x}^* = \sum_{i \in [n]} \alpha_i \boldsymbol{q}_i$. Given that $(\boldsymbol{x}^*)^\top \boldsymbol{B}\boldsymbol{x}^* \leq 1$, we have

$$\sum_{i \in [r]} \alpha_i^2 \lambda_i \leq 1.$$

Hence, the values of $\{\alpha_i\}_{i \in [r]}$ are bounded. On the other hand, let us define a subset $S \subseteq [n]$ of size at most s_1 such that $x_i^* \neq 0$ for each $i \in S$ and $x_j^* = 0$ for each $j \in [n] \setminus S$. Then for each $j \in [n] \setminus S$, we arrive at the following linear system:

$$\sum_{j \in [r+1, n]} \alpha_j \widehat{\boldsymbol{q}}_j = -\sum_{i \in [r]} \alpha_i \widehat{\boldsymbol{q}}_i, \tag{16}$$

where $\widehat{\boldsymbol{q}}_i$ denote a subvector of $\boldsymbol{q}_i$ with indices $[n] \setminus S$ for each $i \in [n]$. For a fixed $\{\alpha_i\}_{i \in [r]}$, since the linear system (16) is nonempty, we let $\bar{\boldsymbol{Q}}\bar{\boldsymbol{\alpha}} = \bar{\boldsymbol{q}}$ denote its minimal linear subsystem such that a submatrix $\bar{\boldsymbol{Q}}$ is non-singular and the index set $\widehat{S}$ of $\bar{\boldsymbol{\alpha}}$ is a subset of $[n] \setminus S$. Thus, we can construct an alternative solution $\widehat{\boldsymbol{\alpha}}$ such that

$$\widehat{\alpha}_i = \begin{cases} \alpha_i, & \text{if } i \in [r], \\ (\bar{\boldsymbol{Q}}^{-1}\bar{\boldsymbol{q}})_i, & \text{if } i \in \widehat{S}, \\ 0, & \text{otherwise,} \end{cases}$$

and $\widehat{\boldsymbol{x}} = \sum_{i \in [n]} \widehat{\alpha}_i \boldsymbol{q}_i$. According to Lemma 1, we have

$$\widehat{\boldsymbol{x}}^\top \boldsymbol{B} \widehat{\boldsymbol{x}} \leq 1, \widehat{\boldsymbol{x}}^\top \boldsymbol{A} \boldsymbol{y}^* = (\boldsymbol{x}^*)^\top \boldsymbol{A} \boldsymbol{y}^*,$$

i.e., $(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*)$ is also optimal to SCCA. Hence,

$$\|\widehat{\boldsymbol{x}}\|_2 \leq \sqrt{\|\bar{\boldsymbol{Q}}^{-1}\bar{\boldsymbol{q}}\|_2^2 + \sum_{i \in [r]} \alpha_i^2}$$

Note that $\sum_{i \in [r]} \alpha_i^2 \leq 1/\lambda_r$ and

$$\|\bar{\boldsymbol{Q}}^{-1}\bar{\boldsymbol{q}}\|_2^2 \leq \|\bar{\boldsymbol{Q}}^{-1}\|_2^2 \|\bar{\boldsymbol{q}}\|_2^2 \leq \frac{1}{s_{\min}(\boldsymbol{B})} \frac{1}{\lambda_r}$$

where $s_{\min}(\boldsymbol{B})$ denotes the smallest nonzero singular values of all the submatrices of $[\boldsymbol{q}_{r+1}, \ldots, \boldsymbol{q}_n]$. In summary, we have

$$\|\widehat{\boldsymbol{x}}\|_2 \leq \sqrt{1/\lambda_r + 1/(\lambda_r s_{\min}(\boldsymbol{B}))}.$$

This completes the proof.   □

A.3   Proof of Proposition 3

Proof. First, for any binary variable $\boldsymbol{z} \in \mathcal{Z}$, suppose $S_1 := \{i : z_i = 1, \forall i \in [n]\}$, $S_2 := \{i - n : z_i = 1, \forall i \in [n+1, n+m]\}$, and $T \subseteq [n+m]$ denotes the support of $\boldsymbol{z}$. Then following the proof of Proposition 1, we can construct a rank-one optimal solution $\boldsymbol{X}^* := \begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^* \\ \boldsymbol{y}^* \end{pmatrix}^\top$ to the maximization problem below that admits the optimal value $\sigma_{\max}\left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right)$, i.e.,

$$\max_{\boldsymbol{X} \in \mathcal{S}_+^{n+m}} \{\operatorname{tr}(\tilde{\boldsymbol{A}}\boldsymbol{X}) : \operatorname{tr}(\tilde{\boldsymbol{B}}\boldsymbol{X}) \leq 1, \operatorname{tr}(\tilde{\boldsymbol{C}}\boldsymbol{X}) \leq 1, X_{ii} = 0, \forall i \in [n+m] \setminus T\}$$

$$= \sigma_{\max}\left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right) \geq f(\boldsymbol{z}),$$

where the inequality is because the maximization problem above relaxes the valid constraints $X_{ii} \leq M_{ii}$ for all $i \in T$ in maximization problem (7). The result in Corollary 1 suggests that $\boldsymbol{x}^*, \boldsymbol{y}^*$ can be bounded and their two norms must not exceed M_1, M_2, which means that the optimal solution $\boldsymbol{X}^*$ satisfies the $X_{ii} \leq M_{ii}$ for all $i \in T$. Therefore, $\boldsymbol{X}^*$ is feasible and optimal to maximization problem (7) and we have that

$$f(\boldsymbol{z}) := \sigma_{\max}\left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right).$$

According to strong duality, the minimization problem (8) admits an optimal value $\sigma_{\max}\left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right)$. Next, we construct its optimal solution $(\theta_1^*, \theta_2^*, \boldsymbol{\lambda}^*)$.

For any given $\epsilon > 0$, we let $\theta_1^* = f(\boldsymbol{z})/2$, $\theta_2^* = f(\boldsymbol{z})/2$, $\widehat{\lambda}_i(\epsilon) = \frac{\epsilon}{M_{ii}|T|}$ for all $i \in T$, and $\widehat{\lambda}_i(\epsilon) = \lambda^*(\epsilon)$ for all $i \in [n] \setminus T$, where

$$\lambda^*(\epsilon) := \left[\lambda_{\max} \left(\boldsymbol{D}_2^\top \left(\boldsymbol{D}_1 + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}_T(\epsilon) \right) \right)^{-1} \boldsymbol{D}_2 - \boldsymbol{D}_3 \right) \right]_+ .$$

It is easy to compute that $\theta_1^* + \theta_2^* + \sum_{i \in [n+m]} \widehat{\lambda}_i(\epsilon) M_{ii} z_i = f(\boldsymbol{z}) + \epsilon$. Thus, for any $\epsilon > 0$, if $(\theta_1^*, \theta_2^*, \widehat{\boldsymbol{\lambda}}(\epsilon))$ were feasible, then it is an ϵ-optimal solution to the minimization problem (8). It remains to verify the feasibility of the solution $(\theta_1^*, \theta_2^*, \widehat{\boldsymbol{\lambda}}(\epsilon))$, i.e., checking the constraint below

$$\begin{pmatrix} \theta_1^* \boldsymbol{B} & -\boldsymbol{A}/2 \\ -\boldsymbol{A}^\top/2 & \theta_2^* \boldsymbol{C} \end{pmatrix} + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}(\epsilon) \right) \succeq 0.$$

By performing the permutation of the rows and columns of the above matrix, it is sufficient to show that the new block matrix

$$\begin{pmatrix} \boldsymbol{D}_1 + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}_T(\epsilon) \right) & \boldsymbol{D}_2 \\ \boldsymbol{D}_2^\top & \boldsymbol{D}_3 + \lambda^*(\epsilon)\boldsymbol{I} \end{pmatrix} \succeq 0, \tag{17}$$

is positive semidefinite.

Since $\begin{pmatrix} \boldsymbol{B}_{S_1,S_1} & -\boldsymbol{A}_{S_1,S_2}/2 \\ -\boldsymbol{A}_{S_1,S_2}^\top/2 & \boldsymbol{C}_{S_2,S_2} \end{pmatrix}$ is a principal submatrix of a positive semidefinite matrix $\begin{pmatrix} \boldsymbol{B} & -\boldsymbol{A}/2 \\ -\boldsymbol{A}^\top/2 & \boldsymbol{C} \end{pmatrix}$, it is also positive semidefinite. According to Lemma 1 and the fact that $\theta_1^* = \theta_2^* = \sigma_{\max}\left(\sqrt{(\boldsymbol{B}_{S_1,S_1})^\dagger} \boldsymbol{A}_{S_1,S_2} \sqrt{(\boldsymbol{C}_{S_2,S_2})^\dagger} \right)/2$, the matrix $\boldsymbol{D}_1$ is also positive semidefinite. As $\epsilon > 0$, the matrix $\boldsymbol{D}_1 + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}_T(\epsilon) \right)$ must be positive definite, which means that

$$\left(\boldsymbol{I} - \left(\boldsymbol{D}_1 + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}_T(\epsilon) \right) \right) \left(\boldsymbol{D}_1 + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}_T(\epsilon) \right) \right)^{-1} \right) \boldsymbol{D}_2 = \boldsymbol{0}.$$

Besides, according to the definition of $\lambda^*(\epsilon)$, we obtain

$$\boldsymbol{D}_3 + \lambda^*(\epsilon)\boldsymbol{I} - \boldsymbol{D}_2^\top \left(\boldsymbol{D}_1 + \mathrm{Diag}\left(\widehat{\boldsymbol{\lambda}}_T(\epsilon) \right) \right)^{-1} \boldsymbol{D}_2 \succeq \boldsymbol{0}.$$

Taking these results together, according to Lemma 1, the constraint in (17) must hold for a given solution $(\theta_1^*, \theta_2^*, \widehat{\boldsymbol{\lambda}}(\epsilon))$. Since the objective value corresponding to $(\theta_1^*, \theta_2^*, \widehat{\boldsymbol{\lambda}}(\epsilon))$ is at most ϵ larger than the optimal value of problem (8), letting $\epsilon \to 0$ and using the closedness of the feasible set in problem (8), we can confirm the optimality of $(\theta_1^*, \theta_2^*, \boldsymbol{\lambda}^*)$ with $\lambda_i^* = 0$ for all $i \in T$ and $\lambda_i^* = \lambda^*$ for all $i \in [n] \setminus T$.

Given the closed-form optimal solution to problem (8), the rest of the proof follows from [23][theorem 7]. $\qquad\square$

A.4   Proof of Theorem 3

Proof. The proof is split into three parts.

Part (i). It suffices to prove that CCA admits an optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying $\|\boldsymbol{x}^*\|_0 \leq r$ and $\|\boldsymbol{y}^*\|_0 \leq \widehat{r}$. Then, $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is also feasible and optimal to SCCA, which implies the equivalence between SCCA and CCA.

First, according to Part (ii) in Proposition 1, we can obtain a closed-form optimal solution $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$ for the CCA. By adjusting $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$, we will construct a new optimal sparse solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfying $\|\boldsymbol{x}^*\|_0 \leq r$ and $\|\boldsymbol{y}^*\|_0 \leq \widehat{r}$.

For matrix $\boldsymbol{B} \in \mathcal{S}_+^n$, we let $\{\boldsymbol{q}_i\}_{i \in [n-r]} \in \mathbb{R}^n$ denote the eigenvectors corresponding to $(n-r)$ zero eigenvalues of $\boldsymbol{B}$. Thus, $\{\boldsymbol{q}_i\}_{i \in [n-r]}$ are orthonormal. There exists a size-$(n-r)$ subset $S \subseteq [n]$ such that the subvectors $\{(\boldsymbol{q}_i)_S\}_{i \in [n-r]}$ are linearly independent, where $(\boldsymbol{q}_i)_S$ denotes the subvector of $\boldsymbol{q}_i$ indexed by S for each $i \in [n-r]$. As a result, there exist a vector $(\gamma_1, \cdots, \gamma_{n-r})^\top$ such that

$$\widehat{\boldsymbol{x}}_S = \sum_{i \in [n]} \gamma_i (\boldsymbol{q}_i)_S. \tag{18}$$

Let us now construct solution $\boldsymbol{x}^*$

$$\boldsymbol{x}^* = \widehat{\boldsymbol{x}} - \sum_{i \in [n-r]} \gamma_i \boldsymbol{q}_i,$$

where $x_i^* = 0$ for all $i \in S$ based on the equation (18) and $|S| = n - r$, implying $\|\boldsymbol{x}^*\|_0 \leq r$. In addition, we show that the new solution $\boldsymbol{x}^*$ is still optimal to CCA. First, $\boldsymbol{x}^*$ is feasible since

$$(\boldsymbol{x}^*)^\top \boldsymbol{B}(\boldsymbol{x}^*) = \widehat{\boldsymbol{x}}^\top \boldsymbol{B}\widehat{\boldsymbol{x}} \leq 1,$$

where the equation is due to $\boldsymbol{B}\boldsymbol{q}_i = \boldsymbol{0}$ for all $i \in [n-r]$.

Given the positive semidefinite block matrix $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} \\ \boldsymbol{A}^\top & \boldsymbol{C} \end{pmatrix}$, using Part (ii) of Lemma 1, the identity $(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^\dagger)\boldsymbol{A} = \boldsymbol{0}$ is equivalent to $\sum_{i \in [n-r]} \boldsymbol{q}_i \boldsymbol{q}_i^\top \boldsymbol{A} = \boldsymbol{0}$. Then, for each $i \in [n-r]$, multiplying $\boldsymbol{q}_i^\top$ on both sides of this equation leads to

$$\boldsymbol{q}_i^\top \left(\sum_{j \in [n-r]} \boldsymbol{q}_j \boldsymbol{q}_j^\top \boldsymbol{A} \right) \boldsymbol{A} = \boldsymbol{q}_i^\top \boldsymbol{0} \Longrightarrow \boldsymbol{q}_i^\top \boldsymbol{A} = \boldsymbol{0},$$

where the result follows from $\boldsymbol{q}_i^\top \boldsymbol{q}_j = 0$ for any $i \neq j$. Then, we can show the optimality of the new solution $\boldsymbol{x}^*$:

$$(\boldsymbol{x}^*)^\top \boldsymbol{A}\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{x}}^\top \boldsymbol{A}\widehat{\boldsymbol{y}} + \sum_{i \in [n-r]} \beta_i \boldsymbol{q}_i^\top \boldsymbol{A}\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{x}}^\top \boldsymbol{A}\widehat{\boldsymbol{y}}.$$

Similarly, we can also construct an optimal sparse solution $\boldsymbol{y}^*$ by leveraging $\widehat{\boldsymbol{y}}$ and eigenvectors of zero eigenvalues of $\boldsymbol{C}$ such that $\|\boldsymbol{y}^*\|_0 \leq s_2$.

Therefore, there exists an optimal solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ to the CCA whose zero norms are bounded from above by $r, \widehat{r}$, respectively. Adding the constraints $\|\boldsymbol{x}\|_0 \leq$

$r, \|\boldsymbol{y}\|_0 \leq \widehat{r}$ to the CCA does not affect the optimality, which gives an equivalent formulation (10) of CCA.

Part (ii). Suppose that $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ denotes an optimal solution to problem (11). When $s_1 \geq r$, following the proof of Part (I), $\tilde{\boldsymbol{x}}$, we can construct another optimal solution $\boldsymbol{x}^*$ whose zero norm is bounded by r and $(\boldsymbol{x}^*, \tilde{\boldsymbol{y}})$ is feasible and optimal to SCCA.

Part (iii). Similarly, we can reduce SCCA to problem (12). We thus complete the proof. □

A.5   Proof of Theorem 4

Proof. Let us first consider the maximization problem over $\boldsymbol{x}$ in (14), i.e.,

$$v_x := \max_{\boldsymbol{x} \in \mathbb{R}^n} \{\boldsymbol{a}^\top \boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x} \leq 1, \|\boldsymbol{x}\|_0 \leq s_1\}. \tag{19}$$

Then, we derive a combinatorial optimization reformulation of problem (19) based on the result below.

Claim 1 *For any subset $S \subseteq [n]$, $\max_{\boldsymbol{x} \in \mathbb{R}^{|S|}} \{\boldsymbol{a}_S^\top \boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{B}_{S,S} \boldsymbol{x} \leq 1\} = \sqrt{\boldsymbol{a}_S^\top (\boldsymbol{B}_{S,S})^\dagger \boldsymbol{a}_S}$.*

Proof. Given $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{b}^\top$, since the matrix $\begin{pmatrix} \boldsymbol{B} & \boldsymbol{a}\boldsymbol{b}^\top \\ \boldsymbol{b}^\top \boldsymbol{a} & \boldsymbol{C} \end{pmatrix}$ is positive semidefinite, using Lemma 1, the identity $(\boldsymbol{I} - \boldsymbol{B}_{S,S}\boldsymbol{B}_{S,S}^\dagger)\boldsymbol{a}_S\boldsymbol{b}^\top = \boldsymbol{0}$ must hold for any subset S. As a result, we have $\boldsymbol{a}_S - \boldsymbol{B}_{S,S}\boldsymbol{B}_{S,S}^\dagger \boldsymbol{a}_S = \boldsymbol{0}$ as vector $\boldsymbol{b}$ is nonzero.

Next, the Lagrangian dual of the problem $\max_{\boldsymbol{x} \in \mathbb{R}^{|S|}} \{\boldsymbol{a}_S^\top \boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{B}_{S,S} \boldsymbol{x} \leq 1\}$ can be written as

$$\max_{\boldsymbol{x} \in \mathbb{R}^{|S|}} \{\boldsymbol{a}_S^\top \boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{B}_{S,S} \boldsymbol{x} \leq 1\} = \min_{\mu \geq 0} \max_{\boldsymbol{x} \in \mathbb{R}^{|S|}} \boldsymbol{a}_S^\top \boldsymbol{x} + \mu - \mu \boldsymbol{x}^\top \boldsymbol{B}_{S,S} \boldsymbol{x}$$

$$= \min_{\mu \geq 0} \mu + \frac{\boldsymbol{a}_S^\top \boldsymbol{B}_{S,S}^\dagger \boldsymbol{a}_S}{4\mu} = \sqrt{\boldsymbol{a}_S^\top (\boldsymbol{B}_{S,S})^\dagger \boldsymbol{a}_S},$$

where the second equation builds on the identity $\boldsymbol{a}_S - \boldsymbol{B}_{S,S}\boldsymbol{B}_{S,S}^\dagger \boldsymbol{a}_S = \boldsymbol{0}$ and optimal solution $\boldsymbol{x}^* = \dfrac{\boldsymbol{B}_{S,S}^\dagger \boldsymbol{a}_S}{\sqrt{\boldsymbol{a}_S^\top (\boldsymbol{B}_{S,S})^\dagger \boldsymbol{a}_S}}$. ◇

Suppose that an optimal solution to problem (19) admits the support S^*. According to Claim 1, we have

$$v_x := \max_{S \subseteq [n], |S| \leq s} \sqrt{\boldsymbol{a}_S^\top (\boldsymbol{B}_{S,S})^\dagger \boldsymbol{a}_S} = \sqrt{\boldsymbol{a}_{S^*}^\top (\boldsymbol{B}_{S^*,S^*})^\dagger \boldsymbol{a}_{S^*}}.$$

On the other hand, the Lagrangian dual of problem (19) can be written as

$$v_x \leq \min_{\lambda \in \mathbb{R}_+} \max_{\boldsymbol{x} \in \mathbb{R}^n} \{\boldsymbol{a}^\top \boldsymbol{x} + \lambda - \lambda \boldsymbol{x}^\top \boldsymbol{B} \boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq s_1\}$$

$$= \min_{\lambda \in \mathbb{R}_+} \max_{S \subseteq [n], |S| \leq s} \lambda + \frac{\boldsymbol{a}_S^\top (\boldsymbol{B}_{S,S})^\dagger \boldsymbol{a}_S}{4\lambda}$$

$$\leq \max_{S\subseteq[n],|S|\leq s} \lambda^* + \frac{\boldsymbol{a}_S^\top (\boldsymbol{B}_{S,S})^\dagger \boldsymbol{a}_S}{4\lambda^*} = \sqrt{\boldsymbol{a}_{S^*}^\top (\boldsymbol{B}_{S^*,S^*})^\dagger \boldsymbol{a}_{S^*}} \leq v_x,$$

where the first equation is due to Claim 1, the second inequality is by plugging the feasible solution $\lambda^* = \frac{\sqrt{\boldsymbol{a}_{S^*}^\top (\boldsymbol{B}_{S^*,S^*})^\dagger \boldsymbol{a}_{S^*}}}{2}$ into minimization problem, and the last equation is from the optimality of subset S^*. Since both left-hand and right-hand sides above equal v_x, the strong duality of problem (19) holds, and all the inequalities above must attain the equalities. That is, problem (19) is equivalent to

$$v_x = \min_{\lambda\in\mathbb{R}_+} \max_{\boldsymbol{x}\in\mathbb{R}^n}\{\boldsymbol{a}^\top\boldsymbol{x} + \lambda - \lambda\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq s_1\}.$$

Since the outer minimization is a one-dimensional convex program that can be solved efficiently, as a result, for any given $\lambda > 0$, the inner maximization is equivalent to solving

$$\max_{\boldsymbol{x}\in\mathbb{R}^n}\{\boldsymbol{a}^\top\boldsymbol{x} - \lambda\boldsymbol{x}^\top\boldsymbol{B}\boldsymbol{x} : \|\boldsymbol{x}\|_0 \leq s_1\}. \tag{20}$$

Next, let us consider the NP-hard sparse regression problem (see, e.g., [28]), which admits

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\{\|\boldsymbol{v} - \boldsymbol{U}\boldsymbol{x}\|_2^2 : \|\boldsymbol{x}\|_0 \leq s\right\} \iff \max_{\boldsymbol{x}\in\mathbb{R}^n} \left\{2\boldsymbol{v}^\top\boldsymbol{U}\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\beta} : \|\boldsymbol{x}\|_0 \leq s\right\},$$
$$\tag{21}$$

where data matrix $\boldsymbol{U}$ consists of observations of n variables and vector $\boldsymbol{v}$ denotes the corresponding response variables.

Suppose that in the problem (20), let us define $\lambda\boldsymbol{B} = \boldsymbol{U}^\top\boldsymbol{U}$ and $\boldsymbol{a} = 2\boldsymbol{U}^\top\boldsymbol{v}$. Then using the singular value decomposition of matrix $\boldsymbol{U}$, we see that the following equation still holds.

$$\boldsymbol{a}_S - \boldsymbol{B}_{S,S}\boldsymbol{B}_{S,S}^\dagger\boldsymbol{a}_S = \boldsymbol{0}, \forall S \subseteq [n].$$

Thus, for any given $\lambda > 0$, the maximization problem (20) is equivalent to the sparse regression problem (21). This shows that problem (19) is NP-hard.

Similarly, the maximization problem over $\boldsymbol{y}$ in (14) can also be reduced to the sparse regression problem. $\qquad\square$

Appendix B: Implementations of greedy and local search algorithms

This section presents the detailed implementations of greedy and local search algorithms based on the formulation (1).

Algorithm 1 Greedy algorithm for SCCA (1)

1: **Input:** Matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathcal{S}_+^m$, $\boldsymbol{C} \in \mathcal{S}_+^m$ and integers $s_1 \in [n]$, $s_2 \in [m]$
2: Compute $(i^*, j^*) \in \text{argmax}_{i \in [m], j \in [n]} \sqrt{(B_{ii})^\dagger} A_{ij} \sqrt{(C_{jj})^\dagger}$
3: Define subsets $\widehat{S}_1 := \{i^*\}$ and $\widehat{S}_2 := \{j^*\}$
4: **for** $\ell = 2, \cdots, \max\{s_1, s_2\}$ **do**
5:     **if** $\ell \leq \min\{s_1, s_2\}$ **then**
6:         $i^* \in \text{argmax}_{i \in [n] \setminus \widehat{S}_1} \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1 \cup \{i\}, \widehat{S}_1 \cup \{i\}})^\dagger} \boldsymbol{A}_{\widehat{S}_1 \cup \{i\}, \widehat{S}_2} \sqrt{(\boldsymbol{C}_{\widehat{S}_2, \widehat{S}_2})^\dagger} \right)$
7:         Update $\widehat{S}_1 := \widehat{S}_1 \cup \{i^*\}$
8:         $j^* \in \text{argmax}_{j \in [m] \setminus \widehat{S}_2} \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1, \widehat{S}_1})^\dagger} \boldsymbol{A}_{\widehat{S}_1, \widehat{S}_2 \cup \{j\}} \sqrt{(\boldsymbol{C}_{\widehat{S}_2 \cup \{j\}, \widehat{S}_2 \cup \{j\}})^\dagger} \right)$
9:     **else if** $s_1 \leq s_2$ **then**
10:         $j^* \in \text{argmax}_{j \in [m] \setminus \widehat{S}_2} \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1, \widehat{S}_1})^\dagger} \boldsymbol{A}_{\widehat{S}_1, \widehat{S}_2 \cup \{j\}} \sqrt{(\boldsymbol{C}_{\widehat{S}_2 \cup \{j\}, \widehat{S}_2 \cup \{j\}})^\dagger} \right)$
11:         Update $\widehat{S}_2 := \widehat{S}_2 \cup \{j^*\}$
12:     **else**
13:         $i^* \in \text{argmax}_{i \in [n] \setminus \widehat{S}_1} \sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1 \cup \{i\}, \widehat{S}_1 \cup \{i\}})^\dagger} \boldsymbol{A}_{\widehat{S}_1 \cup \{i\}, \widehat{S}_2} \sqrt{(\boldsymbol{C}_{\widehat{S}_2, \widehat{S}_2})^\dagger} \right)$
14:         Update $\widehat{S}_1 := \widehat{S}_1 \cup \{i^*\}$
15:     **end if**
16: **end for**
17: **Output:** $\widehat{S}_1, \widehat{S}_2$

Algorithm 2 Local search algorithm for SSVD (1)

1: **Input:** Matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathcal{S}_+^m$, $\boldsymbol{C} \in \mathcal{S}_+^m$ and integers $s_1 \in [n]$, $s_2 \in [m]$
2: Initialize $(\widehat{S}_1, \widehat{S}_2)$ as the output of greedy Algorithm 1
3: **do**
4:     **for** each pair $(i_1, j_1) \in \widehat{S}_1 \times ([n] \setminus \widehat{S}_1)$ **do**
5:         **if** $\sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}, \widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}})^\dagger} \boldsymbol{A}_{\widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}, \widehat{S}_2} \sqrt{(\boldsymbol{C}_{\widehat{S}_2, \widehat{S}_2})^\dagger} \right)$ $>$
        $\sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1, \widehat{S}_1})^\dagger} \boldsymbol{A}_{\widehat{S}_1, \widehat{S}_2} \sqrt{(\boldsymbol{C}_{\widehat{S}_2, \widehat{S}_2})^\dagger} \right)$ **then**
6:             Update $\widehat{S}_1 := \widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}$
7:         **end if**
8:     **end for**
9:     **for** each pair $(i_2, j_2) \in \widehat{S}_2 \times ([m] \setminus \widehat{S}_2)$ **do**
10:         **if** $\sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}, \widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}})^\dagger} \boldsymbol{A}_{\widehat{S}_1 \cup \{j_1\} \setminus \{i_1\}, \widehat{S}_2} \sqrt{(\boldsymbol{C}_{\widehat{S}_2, \widehat{S}_2})^\dagger} \right)$ $>$
         $\sigma_{\max} \left(\sqrt{(\boldsymbol{B}_{\widehat{S}_1, \widehat{S}_1})^\dagger} \boldsymbol{A}_{\widehat{S}_1, \widehat{S}_2} \sqrt{(\boldsymbol{C}_{\widehat{S}_2, \widehat{S}_2})^\dagger} \right)$ **then**
11:             Update $\widehat{S}_2 := \widehat{S}_2 \cup \{j_2\} \setminus \{i_2\}$
12:         **end if**
13:     **end for**
14: **while** there is still an improvement
15: **Output:** $\widehat{S}_1, \widehat{S}_2$
