Appendix: Upper bounds for Model-Free Row-Sparse Principal Component Analysis

Guanyi Wang [1] **Santanu Dey** [1]

1. Problem Setting

The row-sparse principal component analysis problem is defined as follows: Given a sample covariance matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, a sparsity parameter k $(\leq d)$, the task is to find the top-r k-sparsity principal components $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_r) \in \mathbb{R}^{d\times r}$,

$$\underset{\boldsymbol{V}^\top\boldsymbol{V}=\boldsymbol{I}_r,\ \|\boldsymbol{V}\|_0\leq k}{\arg\max} \ \mathrm{Tr}\left(\boldsymbol{V}^\top\boldsymbol{A}\boldsymbol{V}\right). \tag{SPCA}$$

where the row-sparsity constraint $\|\boldsymbol{V}\|_0 \leq k$ denotes that there are at most k non-zero rows in matrix $\boldsymbol{V}$, i.e., the principal components share the global support of cardinality at most k.

1.1. Notations

Let the bold upper case letters, for example, $\boldsymbol{A}, \boldsymbol{B}$ be matrices, and denote its (i, j)-th component as $[\boldsymbol{A}]_{ij}, [\boldsymbol{B}]_{ij}$. Let supp$(\boldsymbol{A})$ be the support of non-zero rows of matrix $\boldsymbol{A}$. Let the bold lower case letters, for example, $\boldsymbol{a}, \boldsymbol{b}$ be vectors, and denote its i-th component as $[\boldsymbol{a}]_i, [\boldsymbol{b}]_i$. Let the regular case letters, for example, I, J be the set of indices. For an integer k, let $[k] := \{1, \dots, k\}$. Given any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $I \subseteq [n], J \subseteq [m]$, let $[\boldsymbol{A}]_{I,J}$ (or $\boldsymbol{A}_{I,J}$ in short) be the sub-matrix of A with rows in I and columns in J. In order to simplify notation, let $[\boldsymbol{A}]_I$ be the sub-matrix of A when considering all rows in I, and let $[\boldsymbol{A}]_j$ be the jth row of matrix $\boldsymbol{A}$. Let the regular lower case letters, for example, α, β be the reals. Let $\oplus$ be the sign of direct plus, i.e., given two square matrices $\boldsymbol{A} \in \mathbb{R}^{n_1\times n_1}, \boldsymbol{B} \in \mathbb{R}^{n_2\times n_2}$, $\boldsymbol{A} \oplus \boldsymbol{B} := \mathrm{diag}(\boldsymbol{A}, \boldsymbol{B}) \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$.

2. Proof of Dual (Upper) Bounds

Let $\mathcal{F}$ be the feasible region of SPCA as

$$\mathcal{F} := \left\{ \boldsymbol{V} \in \mathbb{R}^{d\times r} : \begin{array}{ll} \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_r & (1) \\ \|\boldsymbol{V}\|_0 \leq k & (2) \end{array} \right\},$$

where the constraint (1) is the so-called *Stiefel manifold* (denoted as $\mathrm{St}(d, r)$) and the constraint (2) is the row-sparsity constraint. For the constraint (1), recall that the convex-hull of Stiefel manifold can be explicitly represented (**?**) as,

$$\mathrm{Conv}(\mathrm{St}(d,r)) := \left\{ \boldsymbol{V} : \begin{pmatrix} \boldsymbol{I}_d & -\boldsymbol{V} \\ -\boldsymbol{V}^\top & \boldsymbol{I}_r \end{pmatrix} \succeq \boldsymbol{0}_{d+r} \right\} \Leftrightarrow \{\boldsymbol{V} : \boldsymbol{I}_r - \boldsymbol{V}^\top\boldsymbol{V} \succeq \boldsymbol{0}_r\} \Leftrightarrow \{\boldsymbol{V} : \|\boldsymbol{V}\|_{\mathrm{op}} \leq 1\}.$$

For the constraint (2),

Proposition 2.1. If $\boldsymbol{V} \in \mathcal{F}$, then $\|[\boldsymbol{V}]_{[d],i}\|_1 \leq \sqrt{k}$ holds for all $i \in [r]$.

Proof. Since the operator norm of $\boldsymbol{V}$ is upper bounded by 1, we have that the ℓ_2-norm of each column of $\boldsymbol{V}$ is at most 1. Moreover, each column is k-sparse, then we have $\|[\boldsymbol{V}]_{[d],i}\|_1 \leq \sqrt{k}$ holds for all $i \in [r]$. □

*Equal contribution [1]H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, USA. Correspondence to: Guanyi Wang <gwang93@gatech.edu>, Santanu Dey <santanu.dey@isye.gatech.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

The above proposition can be viewed as the ℓ_1-relaxation of the sparsity constraint for each column in $\boldsymbol{V}$. Moreover, the row-sparsity property can be futher captured by

Proposition 2.2. If $\boldsymbol{V} \in \mathcal{F}$, then $\sum_{j=1}^{d} \|[\boldsymbol{V}]_{j,[r]}\|_2 \leq \sqrt{rk}$.

Proof. For any $\boldsymbol{V} \in \mathcal{F}$, based on the row-sparsity condition $\|\boldsymbol{V}\|_0 \leq k$, there are at most k non-zero values of $\|[\boldsymbol{V}]_{j,[r]}\|_2$ among $j \in [d]$. Since $\|\boldsymbol{V}\|_{\text{op}} \leq 1$ with rank r, then

$$r \geq \|\boldsymbol{V}\|_F^2 = \sum_{j=1}^{d} \|[\boldsymbol{V}]_{j,[r]}\|_2^2,$$

which implies that $\sum_{j=1}^{d} \|[\boldsymbol{V}]_{j,[r]}\|_2 \leq \sqrt{rk}$. □

From Proposition 2.1 and 2.2, we obtain the following result.

Corollary 2.1. [SDP-relaxation] Let $\mathcal{F}$ be the feasible region of SPCA. We have conv$(\mathcal{F})$ is contained in the following convex set

$$\mathcal{C} := \left\{ \boldsymbol{V} : \begin{array}{l} \boldsymbol{I}_r - \boldsymbol{V}^\top \boldsymbol{V} \succeq \boldsymbol{0}_r, \\ \sum_{j=1}^{d} \|\boldsymbol{v}_i\|_1 \leq \sqrt{k}, \ \forall i \in [r] \\ \sum_{j=1}^{d} \|[\boldsymbol{V}]_{j,[r]}\|_2 \leq \sqrt{rk}, \ \forall j \in [d] \end{array} \right\}.$$

Since SDP-relaxations are usually difficult to solve, to be more scalable in practice, instead of using semi-definite constraint, we replace it with second-order cone constraints. In particular, we will replace the constraints defining the convex hull of the Stiefel manifold by a simple second-order-cone representable relaxation to obtain the following result.

Corollary 2.2. [SOCP-relaxation] Let $\mathcal{F}$ be the feasible region of SPCA. We have conv$(\mathcal{F})$ is contained in the following convex set

$$\mathcal{C}' := \left\{ \boldsymbol{V} : \begin{array}{l} \|[\boldsymbol{V}]_{[d],i}\|_2^2 \leq 1, \ \forall i \in [r] \\ \|[\boldsymbol{V}]_{[d],i_1} \pm [\boldsymbol{V}]_{[d],i_2}\|_2^2 \leq 2, \ \forall i_1 \neq i_2 \in [r] \\ \sum_{j=1}^{d} \|[\boldsymbol{V}]_{j,[r]}\|_2 \leq \sqrt{rk} \\ \|[\boldsymbol{V}]_{j,[r]}\|_2 \in [0,1], \ \forall j \in [d] \end{array} \right\}.$$

2.1. Proof of Theorem 1

Let $\text{opt}^{\mathcal{F}}, \text{opt}^{\mathcal{C}}, \text{opt}^{\mathcal{C}'}$ be the optimal values of the following:

$$\text{opt}^{\mathcal{F}} := \max_{\boldsymbol{V} \in \mathcal{F}} \ \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{A} \boldsymbol{V}\right),$$

$$\text{opt}^{\mathcal{C}} := \max_{\boldsymbol{V} \in \mathcal{C}} \ \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{A} \boldsymbol{V}\right), \tag{Relax}$$

$$\text{opt}^{\mathcal{C}'} := \max_{\boldsymbol{V} \in \mathcal{C}'} \ \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{A} \boldsymbol{V}\right). \tag{SOCP-Relax}$$

Our first main result is that:

Theorem 1. $\text{opt}^{\mathcal{F}} \leq \text{opt}^{\mathcal{C}'} \leq (1+\sqrt{r})^2 \text{opt}^{\mathcal{F}}$.

Proof. Consider any $\boldsymbol{V} \in \mathcal{C}'$, sort the row of $\boldsymbol{V}$ by its ℓ_2-norm in decreasing order $\{j_1, j_2, \ldots, j_d\}$, i.e.,

$$\|[\boldsymbol{V}]_{j_1,[r]}\|_2 \geq \ldots \geq \|[\boldsymbol{V}]_{j_d,[r]}\|_2.$$

Decompose the matrix $\boldsymbol{V}$ based on its top-k largest rows, second top-k largest rows, and so on, i.e., let $m = \lceil d/k \rceil$, $\boldsymbol{V} = \boldsymbol{V}^1 + \cdots \boldsymbol{V}^m$ with

$$\text{supp}(\boldsymbol{V}^1) = \{j_1, \ldots, j_k\} =: J^1, \ \ldots, \ \boldsymbol{V}^m = \{j_{d-(m-1)k}, \ldots, j_d\} =: J^m.$$

For each $p = 1, \dots, m$, we have $\|\boldsymbol{V}^p/\|\boldsymbol{V}^p\|_{\text{op}}\|_0 \le k$, $\|\boldsymbol{V}^p/\|\boldsymbol{V}^p\|_{\text{op}}\|_{\text{op}} = 1$, thus $\boldsymbol{V}^p/\|\boldsymbol{V}^p\|_{\text{op}} \in \text{conv}(\mathcal{F})$. Since $\text{Tr}(\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V})$ is convex, then $\max_{\boldsymbol{V}\in\mathcal{F}} \text{Tr}(\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}) = \max_{\boldsymbol{V}\in\text{conv}(\mathcal{F})} \text{Tr}(\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V})$. To verify the approximation ratio,

$$\begin{aligned} &\boldsymbol{V} = \boldsymbol{V}^1 + \cdots \boldsymbol{V}^m = \|\boldsymbol{V}^1\|_{\text{op}} \frac{\boldsymbol{V}^1}{\|\boldsymbol{V}^1\|_{\text{op}}} + \cdots + \|\boldsymbol{V}^m\|_{\text{op}} \frac{\boldsymbol{V}^m}{\|\boldsymbol{V}^m\|_{\text{op}}} \\ \Leftrightarrow\ & \frac{\boldsymbol{V}}{\sum_{p=1}^m \|\boldsymbol{V}^p\|_{\text{op}}} = \frac{\|\boldsymbol{V}^1\|_{\text{op}}}{\sum_{p=1}^m \|\boldsymbol{V}^p\|_{\text{op}}} \frac{\boldsymbol{V}^1}{\|\boldsymbol{V}^1\|_{\text{op}}} + \cdots + \frac{\|\boldsymbol{V}^m\|_{\text{op}}}{\sum_{p=1}^m \|\boldsymbol{V}^p\|_{\text{op}}} \frac{\boldsymbol{V}^m}{\|\boldsymbol{V}^m\|_{\text{op}}} \in \text{conv}(\mathcal{F}). \end{aligned}$$

Notice that

$$\|\boldsymbol{V}^p\|_{\text{op}} \le \max \left\{ 1, \sqrt{\sum_{j\in J^p} \|[\boldsymbol{V}]_{j,[r]}\|_2^2} \right\},$$

then based on the decomposition of ℓ_2 norms of rows,

$$\begin{aligned} \sum_{p=1}^m \|\boldsymbol{V}^p\|_{\text{op}} &= \|\boldsymbol{V}^1\|_{\text{op}} + \sum_{p=2}^m \|\boldsymbol{V}^p\|_{\text{op}} \\ &\le 1 + \sum_{p=2}^m \sqrt{\left(\frac{\sum_{j\in J^{p-1}} \|[\boldsymbol{V}]_{j,[r]}\|_2}{k} \right)^2 \cdot k} \\ &\le 1 + \frac{1}{\sqrt{k}} \cdot \sum_{p=2}^m \sum_{j\in J^{p-1}} \|[\boldsymbol{V}]_{j,[r]}\|_2 \\ &\le 1 + \frac{1}{\sqrt{k}} \sum_{j=1}^d \|[\boldsymbol{V}]_{j,[r]}\|_2 \\ &\le 1 + \sqrt{r} \end{aligned}$$

where the final inequality holds since the constraint $\sum_{j=1}^d \|[\boldsymbol{V}]_{j,[r]}\|_2 \le \sqrt{rk}$ in $\mathcal{C}'$. Therefore, we have

$$\boldsymbol{V} \in \left(\sum_{p=1}^m \|\boldsymbol{V}^p\|_{\text{op}} \right) \cdot \text{conv}(\mathcal{F}) \subseteq (1+\sqrt{r}) \cdot \text{conv}(\mathcal{F}),$$

i.e., $\mathcal{C}' \subseteq (1+\sqrt{r}) \cdot \text{conv}(\mathcal{F})$. Hence $\text{opt}^{\mathcal{F}} \le \text{opt}^{\mathcal{C}'} \le (1+\sqrt{r})^2 \text{opt}^{\mathcal{F}}$ holds. □

A corollary can be derived from the Theorem 1 based on the containment $\mathcal{C} \subseteq \mathcal{C}'$ as follows:

Corollary 2.3. $\text{opt}^{\mathcal{F}} \le \text{opt}^{\mathcal{C}} \le (1+\sqrt{r})^2 \text{opt}^{\mathcal{F}}$.

Remark: *For $r = 1$ case, Theorem 1 and Corollary 2.3 provide constant multiplicative approximation ratios. Thus inapproximability results from (**??**) implies that solving Relax or SOCP-Relax to optimality is NP-hard.*

2.2. Proof of Proposition 2.3 & 2.4

To overcome the non-convex part of Relax or SOCP-Relax, the objective function of Relax or SOCP-Relax is further relaxed via piecewise-linear functions using special-ordered sets type-2 constraints. Recall that the piecewise-linear approximation (PLA) set is defined as

$$\text{PLA} := \left\{ (g, \xi, \eta) : \begin{array}{l} g_{ji} = \boldsymbol{a}_j^\top \boldsymbol{v}_i,\ (j,i) \in [d]\times[r] \\ g_{ji} = \sum_{\ell=-N}^N \gamma_{ji}^\ell \eta_{ji}^\ell \\ \xi_{ji} = \sum_{\ell=-N}^N \left(\gamma_{ji}^\ell\right)^2 \eta_{ji}^\ell \\ \left(\eta_{ji}^\ell\right)_{\ell=-N}^N \in \text{SOS-II} \end{array} \right\}$$

in which SOS-II denotes the set of special-ordered sets type-2 constraints as follows:

$$\text{SOS-II} := \left\{ \left(\eta_{ji}^{\ell}\right)_{\ell=-N}^{N} : \begin{array}{ll} \sum_{\ell=-N}^{N} \eta_{ji}^{\ell} = 1 & \\ \sum_{\ell=-N}^{N-1} y^{\ell} = 1 & \\ \eta_{ji}^{\ell} + \eta_{ji}^{\ell+1} \leq y^{\ell} & \ell = -N, \ldots, N-1 \\ \eta_{ji}^{\ell} \geq 0 & \ell = -N, \ldots, N \\ y^{\ell} \in \{0,1\} & \ell = -N, \ldots, N-1 \end{array} \right\}$$

since SOS-II contains the integer variables, we name the convex relaxation of Relax or SOCP-Relax be *semi-definite convex integer program* (SDCIP)

$$\begin{array}{rl} \text{ub}^{\mathcal{C}} := \max & \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} \xi_{ji} \\ \text{s.t.} & \boldsymbol{V} \in \mathcal{C} \\ & (g, \xi, \eta) \in \text{PLA} \end{array} \tag{SDCIP}$$

or *second-order-cone convex integer program* (SOCIP)

$$\begin{array}{rl} \text{ub}^{\mathcal{C}'} := \max & \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} \xi_{ji} \\ \text{s.t.} & \boldsymbol{V} \in \mathcal{C}' \\ & (g, \xi, \eta) \in \text{PLA} \end{array} \tag{SOCIP}$$

Thus we have the Proposition 2.3 and Proposition 2.4,

Proposition 2.3. The optimal value $\text{ub}^{\mathcal{C}'}$ of SOCIP is an upper bound of SPCA.

Proof. Based on Theorem 2.3, we have the optimal value $\text{opt}^{\mathcal{F}}$ of SPCA

$$\text{opt}^{\mathcal{F}} := \max_{\boldsymbol{V}^{\top}\boldsymbol{V}=\boldsymbol{I}_r, \|\boldsymbol{V}\|_0 \leq k} \text{Tr}\left(\boldsymbol{V}^{\top}\boldsymbol{A}\boldsymbol{V}\right)$$

is upper bounded by the optimal value $\text{opt}^{\mathcal{C}'}$ of Relax

$$\text{opt}^{\mathcal{C}} := \max_{\boldsymbol{V} \in \mathcal{C}'} \text{Tr}\left(\boldsymbol{V}^{\top}\boldsymbol{A}\boldsymbol{V}\right).$$

To show that $\text{ub}^{\mathcal{C}'}$ is an upper bound, it is sufficient to show that $\text{ub}^{\mathcal{C}'} \geq \text{opt}^{\mathcal{C}'}$. Consider the auxiliary variable ξ_{ji} for all $(j,i) \in [d] \times [r]$ with $g_{ji} = \boldsymbol{a}_j^{\top}\boldsymbol{v}_i$. Based on the property of SOS-II constraint, for a fixed (j,i), there are at most two non-zero continuous variables in η_{ji}^{ℓ}, say $\eta_{ji}^{\ell^*}, \eta_{ji}^{\ell^*+1}$, such that $\eta_{ji}^{\ell^*} + \eta_{ji}^{\ell^*+1} = 1$. Combining with the constraints in set PLA, we have

$$g_{ji} = \gamma_{ji}^{\ell^*}\eta_{ji}^{\ell^*} + \gamma_{ji}^{\ell^*+1}\eta_{ji}^{\ell^*+1}, \qquad \xi_{ji} = \left(\gamma_{ji}^{\ell^*}\right)^2 \eta_{ji}^{\ell^*} + \left(\gamma_{ji}^{\ell^*+1}\right)^2 \eta_{ji}^{\ell^*+1},$$

and hence,

$$\xi_{ji} - g_{ji}^2 = \left(\gamma_{ji}^{\ell^*+1} - \gamma_{ji}^{\ell^*}\right)^2 \eta_{ji}^{\ell^*}\eta_{ji}^{\ell^*+1} \geq 0.$$

That is to say, the objective function $\sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} \xi_{ji}$ in SOCIP is greater than or equal to the function $\sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} g_{ji}^2$, where $\sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} g_{ji}^2$ is equivalent to the objective function $\text{Tr}\left(\boldsymbol{V}^{\top}\boldsymbol{A}\boldsymbol{V}\right)$ in Relax by the definition of g_{ji}. Therefore, $\text{ub}^{\mathcal{C}'} \geq \text{opt}^{\mathcal{C}'}$ holds. $\square$

Proposition 2.4. The optimal value $\text{ub}^{\mathcal{C}'}$ of SOCIP can be upper bounded by $(1+\sqrt{r})^2\text{opt}^{\mathcal{F}} + \sum_{j=1}^{d} \frac{r\lambda_j\theta_j^2}{4N^2}$, which is an affine function of $\text{opt}^{\mathcal{F}}$.

Proof. For SOCIP, the objective function $\text{Tr}\left(\boldsymbol{V}^{\top}\boldsymbol{A}\boldsymbol{V}\right)$ equals to $\sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} (\boldsymbol{a}_j^{\top}\boldsymbol{v}_i)^2 = \sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} g_{ji}^2$. By Theorem 1, we have $\sum_{j=1}^{d} \lambda_j \sum_{i=1}^{r} g_{ji}^2 \leq 4r\text{opt}^{\mathcal{F}}$. Since $g_{ji} \in [-\theta_j, \theta_j]$, without any prior information, we split the interval $[-\theta_j, \theta_j]$ evenly via splitting points $(\gamma_{ji}^\ell)_{\ell=-N}^N$ such that $\gamma_{ji}^\ell = \frac{\ell}{N} \cdot \theta_j$. Based on the proof of Proposition 2.1, we have

$$\xi_{ji} - g_{ji}^2 = \left(\gamma_{ji}^{\ell^*+1} - \gamma_{ji}^{\ell^*}\right)^2 \eta_{ji}^{\ell^*} \eta_{ji}^{\ell^*+1} = \frac{\theta_j^2}{N^2} \eta_{ji}^{\ell^*} \eta_{ji}^{\ell^*+1} \leq \frac{\theta_j^2}{4N^2}.$$

Therefore, the objective function in SOCIP $\sum_{j=1}^d \lambda_j \sum_{i=1}^r \xi_{ji} \leq \sum_{j=1}^d \lambda_j \sum_{i=1}^r g_{ji}^2 + \sum_{j=1}^d \frac{r\lambda_j \theta_j^2}{4N^2} \leq (1+\sqrt{r})^2 \text{opt}^{\mathcal{F}} + \sum_{j=1}^d \frac{r\lambda_j\theta_j^2}{4N^2}$. □

3. Proof of Primal (Lower) Bounds

Recall that Given a sample covariance matrix $\boldsymbol{A}$, let $\boldsymbol{A}^{1/2}$ be its positive semi-definite square root such that $\boldsymbol{A} = \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}$, the SPCA can be represented in the following fashion:

$$\begin{aligned} \min_{\boldsymbol{V} \in \mathbb{R}^{d\times r}} \quad & \left\|\boldsymbol{A}^{1/2} - \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{A}^{1/2}\right\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_r \\ & \|\boldsymbol{V}\|_0 \leq k \end{aligned} \tag{SPCA-lasso}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Furthermore, SPCA-lasso can be reformulated into the following *two-stage (inner & outer) optimization problem*:

$$\underbrace{\min_{S\subseteq[d],\ |S|\leq k}}_{\text{outer optimization}} \ \underbrace{\min_{[\boldsymbol{V}]_S^\top[\boldsymbol{V}]_S = \boldsymbol{I}_r}}_{\text{inner optimization}} \|[\boldsymbol{A}^{1/2}]_S - [\boldsymbol{V}]_S[\boldsymbol{V}]_S^\top[\boldsymbol{A}^{1/2}]_S\|_F^2 + \|[\boldsymbol{A}^{1/2}]_{S^C}\|_F^2$$

Given support S, there is a closed form solution of the inner optimization by eigenvalue decomposition. Thus the main challenging of solving SPCA-lasso is to find a support set S within $\binom{d}{k}$ possible support set, which is known to be NP-hard. The local search algorithm 1 is therefore proposed to find a relative good primal solution (via updating the support set S in each epoch).

Algorithm 1 Local Search Method

Input: Covariance matrix $\boldsymbol{A}$, sparsity parameter k, number of eigenvectors r, number of maximum iterations T.
Output: A feasible solution $\boldsymbol{V}$ for SPCA.

```
Initialize with S_0 ⊆ [d] with |S_0| = k.
while epoch t = 1, ..., T do
  For each j ∈ S_{t-1}, set the reduced value Δ_j, set removing candidate j^out := arg min_{j∈S_{t-1}} Δ_j.
  For each j' ∈ S^C_{t-1}, set the will-reduced value Δ_{j'}, set adding candidate j^in := arg min_{j'∈S_{t-1}} Δ_{j'}
  if Δ_{j^in} > Δ_{j^out} then
    Set S_t := S_{t-1} − {j^out} + {j^in}.
    By eigenvalue decomposition, [A^{1/2}]_{S_t}[A^{1/2}]^⊤_{S_t} = U_{S_t} Λ_{S_t} U^⊤_{S_t}, set [V]_{S_t} = [U_{S_t}]_{[k],[r]}.
  else
    Break while loop.
  end if
end while
Return V.
```

Remark 3.1. Note that, for general instance (i.e., model-free case), we initialize the support set S_0 by picking $S_0 \subseteq [d]$ uniformly at random, and repeat the Algorithm 1 for serval times for best solution. But in Section 3.2, we show that under some statistical assumptions, a specific initialization method will find the support set with respect to the optimal solution with high probability.

3.1. Primal Method in Model-free Case

Here we arrive the following results: In model-free case,

Theorem 2. Algorithm 1 is a monotone decreasing algorithm in the objective value of SPCA.

Proof. Based on the updating rule in Algorithm 1, when $\Delta_{j^{\text{in}}} > \Delta_{j^{\text{out}}}$ and set $S_t = S_{t-1} - \{j^{\text{out}}\} + \{j^{\text{in}}\}$, the objective value of SPCA-lasso satisfies

$$\begin{aligned}\left\|\boldsymbol{A}^{1/2} - \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{A}^{1/2}\right\|_F^2 &= \left\|[\boldsymbol{A}^{1/2}]_{S_{t-1}} - [\boldsymbol{V}]_{S_{t-1}}[\boldsymbol{V}]_{S_{t-1}}^\top [\boldsymbol{A}^{1/2}]_{S_{t-1}}\right\|_F^2 + \left\|[\boldsymbol{A}^{1/2}]_{S_{t-1}^C}\right\|_F^2 \\ &> \left\|[\boldsymbol{A}^{1/2}]_{S_t} - [\boldsymbol{V}]_{S_{t-1}}[\boldsymbol{V}]_{S_{t-1}}^\top [\boldsymbol{A}^{1/2}]_{S_t}\right\|_F^2 + \left\|[\boldsymbol{A}^{1/2}]_{S_t^C}\right\|_F^2 \\ &\geq \left\|[\boldsymbol{A}^{1/2}]_{S_t} - [\boldsymbol{V}]_{S_t}[\boldsymbol{V}]_{S_t}^\top [\boldsymbol{A}^{1/2}]_{S_t}\right\|_F^2 + \left\|[\boldsymbol{A}^{1/2}]_{S_t^C}\right\|_F^2 .\end{aligned}$$

Thus Algorithm 1 is a monotone decreasing algorithm. □

Theorem 3. Algorithm 1 terminates in at most $\binom{d}{k}$ epochs.

Proof. We claim that if a subset $S \subseteq [d]$ exists in tth epoch, then such set S will not exist in future epochs. Otherwise, suppose $S_{t_1} = S_{t_2}$ and $t_1 < t_2$, then based on the monotonicity of Algorithm 1, we have

$$\begin{aligned}&\left\|[\boldsymbol{A}^{1/2}]_{S_{t_1}} - [\boldsymbol{V}]_{S_{t_1}}[\boldsymbol{V}]_{S_{t_1}}^\top [\boldsymbol{A}^{1/2}]_{S_{t_1}}\right\|_F^2 + \left\|[\boldsymbol{A}^{1/2}]_{S_{t_1}^C}\right\|_F^2 \\ > &\left\|[\boldsymbol{A}^{1/2}]_{S_{t_2}} - [\boldsymbol{V}]_{S_{t_2}}[\boldsymbol{V}]_{S_{t_2}}^\top [\boldsymbol{A}^{1/2}]_{S_{t_2}}\right\|_F^2 + \left\|[\boldsymbol{A}^{1/2}]_{S_{t_2}^C}\right\|_F^2\end{aligned}$$

which contradicts to $S_{t_1} = S_{t_2}$, and the claim holds. Since there are at most $\binom{d}{k}$ possible subsets, then the Algorithm 1 terminates in finite epochs. □

3.2. Primal Method with Statistical Model

Although this is not the main contribution of our paper, to complete the framework, we demonstrate that the primal feasible solution obtained from our primal heuristic ensures some statistical properties when the following assumptions hold.

Assumption 1. Assume that $\{\boldsymbol{x}_m\}_{m=1}^M \in \mathbb{R}^d$ is a sequence of i.i.d. random samples generated from Gaussian distribution with zero mean and *true covariance matrix* $\boldsymbol{\Sigma}$.

To demonstrate the main difference between usual sparse PCA and row-sparse PCA defined in SPCA, consider the following *block spiked covariance matrix model* (described in Assumption 2) and *support-inconsistency condition* (Assumption 3).

Assumption 2. Let Θ be the collection of all true covariance matrix $\boldsymbol{\Sigma}$ with block diagonal structure $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 \oplus \boldsymbol{\Sigma}_2 \oplus \boldsymbol{0}_{\text{rest}}$ such that:

- $\boldsymbol{\Sigma}_1 \in \mathbb{R}^{k\times k}$ is a rank $r^*(< k)$ matrix with top-r^* largest eigen-pairs $(\lambda_1, \boldsymbol{v}_1), \ldots, (\lambda_{r^*}, \boldsymbol{v}_{r^*})$, $\boldsymbol{\Sigma}_2 \in \mathbb{R}^{k\times k}$ is a full-rank matrix with the next top-k largest eigen-pairs $(\lambda_{r^*+1}, \boldsymbol{v}_{r^*+1}), \ldots, (\lambda_{r^*+k}, \boldsymbol{v}_{r^*+k})$, in which $\lambda_1 > \cdots > \lambda_{r^*} > \lambda_{r^*+1} > \cdots > \lambda_{r^*+k} > 0$.
- The diagonal entries of $\boldsymbol{\Sigma}$ satisfy $\max_{j\in[k]}[\boldsymbol{\Sigma}_1]_{j,j} < \min_{j\in[k]}[\boldsymbol{\Sigma}_2]_{j,j}$, i.e., the top-$k$ diagonal entries are all in sub-matrix $\boldsymbol{\Sigma}_2$.

Let $\boldsymbol{\Sigma}$ be a true covariance matrix in set Θ.

Assumption 3. Assume that $\boldsymbol{\Sigma}$ satisfies: the sum of eigenvalues in first block $\boldsymbol{\Sigma}_1$ is upper bounded by $\sum_{j=r^*+1}^{2r^*+1} \lambda_j$, i.e.,

$$\sum_{j=1}^{r^*} \lambda_j < \sum_{j=r^*+1}^{2r^*+1} \lambda_j .$$

Remark 3.2. Based on Assumption 2 and 3, easy to observe that:

- given k and $r \leq r^*$, the global optimal solution of SPCA with truth covariance matrix $\mathbf{\Sigma}$ has support set $\{1, \dots, k\}$,
- in contrast, given k and $r > r^*$, the global optimal solution of SPCA with truth covariance matrix $\mathbf{\Sigma}$ has support set $\{k+1, \dots, 2k\}$.

Thus a significant shortage of existing support recovery methods is that these methods fail to recover the optimal support set of SPCA.

To overcome this shortage, we proposed an initialization method that recovers the optimal support set with two procedures:

- **Recover the union of support sets:** Based on the existing results in **??**, the *Covariance Thresholding* algorithm (Algorithm 1 in **?**) with input sample covariance matrix $\boldsymbol{A} = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{x}_m \boldsymbol{x}_m^\top$, sparse parameter $2k$, thresholding parameter τ, ρ (defined in Theorem 1, Theorem 3 of **?**) is able to recover the *union of support sets*

  $$S := \bigcup_{i=1}^{r^*+k} \text{supp}(\boldsymbol{v}_i)$$

  using the *soft-thresholding matrix* $\hat{\boldsymbol{A}}$ (mentioned in Algorithm 1 in **?**) with high probability, if the following Assumption 4 is satisfied.

  Assumption 4. Let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_d$ be the set of eigenvectors of $\mathbf{\Sigma}$ corresponding to eigenvalues $\lambda_1, \dots, \lambda_d$ respectively as we defined in Assumption 2. Let $S_1, \dots, S_d$ be the set of supports of $\boldsymbol{v}_1, \dots, \boldsymbol{v}_d$. There exists constants $c_1, c_2 > 0$ such that the following holds. The non-zero components satisfy $|[\boldsymbol{v}_j]_i| \geq c_1/\sqrt{k}$ for all $i \in S_j$ and $j = 1, \dots, r^*+k$. Furthermore, for any j, j', $|[\boldsymbol{v}_j]_i/[\boldsymbol{v}_{j'}]_i| \leq c_2$ for all $i \in S_j \cap S_{j'}$. Here we assume that $c_2 \geq 1$.

  In the block spiked covariance matrix assumption, the union of support sets is the set with respect to the blocks $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$.

- **Search for optimal support for SPCA:** In the second procedure, we search for the global optimal support set of SPCA problem based the structure of blocked spiked covariance matrix.

Here is the pseudocode of Initialization method 2.

Algorithm 2 Initialization - Statistical Model

Input: Sample covariance matrix $\boldsymbol{A}$, sparsity parameter k, number of eigenvectors r.
Output: An initial support set S_0.

Do Covariance Thresholding Algorithm with input $(\boldsymbol{A}, 2k, \tau, \rho)$ described in **?**, let S be the union of support sets, and $\hat{\boldsymbol{A}}$ be the soft-thresholding matrix of $\boldsymbol{A}$.
Set $\bar{\boldsymbol{A}} = [\hat{\boldsymbol{A}}]_{S,S}$. Without loss of generality, let $\{k+1, \dots, 2k\}$ be the set of indices corresponding to the top-k diagonal entries in $\bar{\boldsymbol{A}}$, and let $\{1, \dots, k\}$ be the rest.
Return support set $S_0 = \{1, \dots, k\}$ if $r \leq r^*$, and $S_0 = \{k+1, \dots, 2k\}$ if $r > r^*$.

Notice that Assumptions 1, 2, 3, 4 can hold simultaneously, one possible example is the artificial instance mentioned in Section 3 of the main content.

Let $\beta := \min_{q \neq q'}\{\lambda_1 - \lambda_2, \lambda_2 - \lambda_3, \dots, \lambda_{r^*+k-1} - \lambda_{r^*+k}, \lambda_{r^*+k}\}$ be the *minimum eigenvalue gap*.

Theorem 4. Assume the assumptions 1, 2, 3, 4 hold, and further we have $M \geq \frac{C^2}{\epsilon^2}(2k)^2 \max\{\beta, 1\} \log \frac{d}{(2k)^2}, M > (2k)^2, \frac{d}{e} > (2k)^2$ for C a constant. If

$$\max_{j \in [k]} [\mathbf{\Sigma}_1]_{j,j} + \epsilon < \min_{j \in [k]} [\mathbf{\Sigma}_2]_{j,j}$$

then the solution $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ obtained from Algorithm 1 satisfies

$$\text{Tr}(\boldsymbol{V}^\top \mathbf{\Sigma} \boldsymbol{V}) \geq \max_{\boldsymbol{V} \in \mathcal{F}} \text{Tr}(\boldsymbol{V}^\top \mathbf{\Sigma} \boldsymbol{V}) - 2r\epsilon$$

with high probability $1 - o(1)$.

Proof. We show this in the following steps:

Estimating the sample covariance matrix $\boldsymbol{A}$: Recall the Theorem 1 and Covariance Thresholding Algorithm in **?**, there exists a constant $C > 0$ such that: for any $\epsilon > 0$, when the number of samples $M \geq \frac{C^2}{\epsilon^2}(2k)^2 \max\{\beta, 1\} \log \frac{d}{(2k)^2}$, the *soft-thresholding matrix* $\hat{\boldsymbol{A}}$ (mentioned in Covariance Thresholding Algorithm) satisfies $\|\hat{\boldsymbol{A}} - \boldsymbol{\Sigma}\|_{\text{op}} \leq \epsilon$ with probability $1 - o(1)$. This implies that for any diagonal entries $\hat{\boldsymbol{A}}_{jj}$ with $j = 1, \ldots, d$, we have $|\hat{\boldsymbol{A}}_{jj} - \hat{\boldsymbol{\Sigma}}_{jj}| < \epsilon$ holds with probability $1 - o(1)$.

Recovering the union of supports: Since the Assumptions 1, 2, 3, 4 hold, and the sample size M satisfies the conditions given in 4, Theorem 3 in (**?**) guarantees that: there exists a constant $C_0 = C_0(c_1, c_2, \lambda_1, \beta)$ such that if $M \geq C_0(2k)^2(k + r^*) \log \frac{d}{(2k)^2}$, then the covariance thresholding algorithm recovers the union of supports $\bigcup_{j=1}^{r^*+k} S_j$ with probability $1 - o(1)$.

Recovering the exact support of SPCA: Note that the set S obtained from the Covariance Thresholding Algorithm only recovers the union of supports $\bigcup_{j=1}^{r^*+k} S_j$ with high probability. Condition on S is successfully recovered, suppose $\max_{j \in [k]}[\boldsymbol{\Sigma}_1]_{j,j} + 2\epsilon < \min_{j \in [k]}[\boldsymbol{\Sigma}_2]_{j,j}$, we have the block indexed by the top-k diagonal entries corresponding to the block $\boldsymbol{\Sigma}_2$ since $|\hat{\boldsymbol{A}}_{jj} - \hat{\boldsymbol{\Sigma}}_{jj}| < \epsilon$ holds with probability $1 - o(1)$. Hence the initialization method 2 recover the true support by S_0.

Quality of primal feasible solution: Let $\boldsymbol{V}$ be the primal solution obtained from Algorithm 1. Let $\boldsymbol{A}_{S_0,S_0} = \boldsymbol{U}_{S_0}\boldsymbol{\Lambda}_{S_0}\boldsymbol{U}_{S_0}^\top \in \mathbb{R}^{k \times k}$ be the eigenvalue decomposition of $\boldsymbol{A}_{S_0,S_0}$ with diagonal entries in $\boldsymbol{\Lambda}_{S_0}$ sorted in decreasing order. We have the solution obtained by Algorithm 1 satisfies:

$$
\begin{aligned}
\text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{\Sigma} \boldsymbol{V}\right) &\geq \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{A} \boldsymbol{V}\right) - \left|\text{Tr}\left(\boldsymbol{V}^\top (\boldsymbol{\Sigma} - \boldsymbol{A}) \boldsymbol{V}\right)\right| \\
&\geq \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{A} \boldsymbol{V}\right) - r\epsilon && \text{by } \|\hat{\boldsymbol{A}} - \boldsymbol{\Sigma}\|_{\text{op}} \leq \epsilon \\
&\geq \text{Tr}\left([\boldsymbol{U}_{S_0}]_{[k],[r]}^\top \boldsymbol{A}_{S_0,S_0} [\boldsymbol{U}_{S_0}]_{[k],[r]}\right) - r\epsilon && \text{by monotone local search 1 with initial } [\boldsymbol{U}_{S_0}]_{[k],[r]} \\
&= \sum_{i=1}^{r} \lambda_i(\boldsymbol{A}_{S_0,S_0}) - r\epsilon && \text{let } \lambda_i(\boldsymbol{A}_{S_0,S_0}) \text{ be } i\text{-th largest eigenvalue} \\
&\geq \sum_{i=1}^{r} \lambda_i(\boldsymbol{\Sigma}_{S_0,S_0}) - 2r\epsilon && \text{by } \|\hat{\boldsymbol{A}} - \boldsymbol{\Sigma}\|_{\text{op}} \leq \epsilon \\
&= \max_{\boldsymbol{V} \in \mathcal{F}} \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{\Sigma} \boldsymbol{V}\right) - 2r\epsilon
\end{aligned}
$$

with probability $1 - o(1)$.

Therefore, we have the quality of primal feasible solution $\boldsymbol{V}$ is close enough to the global optimal value. □

3.3. Proof of Proposition 1.1

However, the above Theorem 4 requires the Assumptions 1, 2, 3, 4. In a more general case, suppose the support set cannot be recovered correctly. We could still verify the quality of a given primal feasible solution by solving SOCIP-impl. Here we formally state the Proposition 1.1:

Proposition 1.1. Let samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ be i.i.d. generated from sub-Gaussian distribution with zero-mean, and $\boldsymbol{\Sigma}$ as the *true underlying covariance matrix (second moment)*. Let $\boldsymbol{A} := \frac{1}{M}\boldsymbol{X}\boldsymbol{X}^\top$ be the sample covariance matrix defined as before. Let $\boldsymbol{V}_{\text{app}}$ be a $(1 - \Delta)$-approximation primal feasible solution corresponding to $\boldsymbol{A}$ where the value of Δ is obtained from its upper (dual) bound via solving SOCIP-impl. If the number of samples M is sufficiently large (i.e., $M \geq \left[(C\sqrt{d} + t)/\epsilon\right]^2$ with C a constant depends on the sub-Gaussian norm of $\boldsymbol{x}$ for any $t > 0$), then:

$$
\text{Tr}\left(\boldsymbol{V}_{\text{app}}^\top \boldsymbol{\Sigma} \boldsymbol{V}_{\text{app}}\right) \geq (1 - \Delta) \cdot \max_{\boldsymbol{V} \in \mathcal{F}} \text{Tr}\left(\boldsymbol{V}^\top \boldsymbol{\Sigma} \boldsymbol{V}\right) - (2 - \Delta) r\epsilon
$$

holds with high probability, where $\epsilon > 0$ is any given constant.

Proof. Note the Remark 5.40 in **?** ensures that: as $M \geq \left[(C\sqrt{d}+t)/\epsilon\right]^2$ with C and t be constants defined as above, we have $\|\boldsymbol{A}-\boldsymbol{\Sigma}\|_{\text{op}} \leq \epsilon$ holds with probability at least $1-2\exp(-ct^2)$ with $c>0$ another constant only depends on the sub-Gaussian norm of $\boldsymbol{x}$.

Now we have the objective value corresponding to $\boldsymbol{V}_{\text{app}}$ satisfies

$$
\begin{aligned}
&\operatorname{Tr}\left(\boldsymbol{V}_{\text{app}}^\top \boldsymbol{\Sigma} \boldsymbol{V}_{\text{app}}\right) \\
\geq\ &\operatorname{Tr}\left(\boldsymbol{V}_{\text{app}}^\top \boldsymbol{A} \boldsymbol{V}_{\text{app}}\right) - r\epsilon && \text{by } \|\boldsymbol{A}-\boldsymbol{\Sigma}\|_{\text{op}} \leq \epsilon \text{ with high probability} \\
\geq\ &(1-\Delta)\cdot \max_{\boldsymbol{V}\in\mathcal{F}} \operatorname{Tr}\left(\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}\right) - r\epsilon && \text{by definition of } (1-\Delta)\text{-approximation primal feasible} \\
\geq\ &(1-\Delta)\cdot \max_{\boldsymbol{V}\in\mathcal{F}} \operatorname{Tr}\left(\boldsymbol{V}^\top \boldsymbol{\Sigma}\boldsymbol{V}\right) - (2-\Delta)r\epsilon && \text{by } \|\boldsymbol{A}-\boldsymbol{\Sigma}\|_{\text{op}} \leq \epsilon \text{ with high probability.}
\end{aligned}
$$

Therefore, the proposition 1.1 holds with high probability. □

Remark 3.3. Notice that the Proposition 1.1 does not require Assumptions 1, 2, 3, 4. In contrast, since the gap between the primal bounds and the dual bounds of $\boldsymbol{A}$ is computed via convex integer program SOCIP-impl, the Proposition 1.1 only request a generalized version of Assumption 1 to ensure the a "good" estimating for the true covariance $\boldsymbol{\Sigma}$.

4. Polynomial Running Time

Recall that the second-order-cone $\mathcal{C}'$ and revised piecewise linear upper approximation set PLA$'$ are defined as follows:

$$
\mathcal{C}' := \left\{ \boldsymbol{V} : \begin{array}{l} \|[\boldsymbol{V}]_{[d],i}\|_2^2 \leq 1,\ \forall i \in [r] \\ \|[\boldsymbol{V}]_{[d],i_1} \pm [\boldsymbol{V}]_{[d],i_2}\|_2^2 \leq 2,\ \forall i_1 \neq i_2 \in [r] \\ \sum_{j=1}^d \|[\boldsymbol{V}]_{j,[r]}\|_2 \leq \sqrt{rk} \\ \|[\boldsymbol{V}]_{j,[r]}\|_2 \in [0,1],\ \forall j \in [d] \end{array} \right\}.
$$

$$
\text{PLA}' := \left\{ (g,\xi,\eta) : \begin{array}{l} g_{ji} = \boldsymbol{a}_j^\top \boldsymbol{v}_i,\ (j,i) \in [d]\times[r] \\ g_{ji} = \sum_{\ell=-N}^N \gamma_{ji}^\ell \eta_{ji}^\ell,\ (j,i) \in J^+ \times [r] \\ \xi_{ji} = \sum_{\ell=-N}^N \left(\gamma_{ji}^\ell\right)^2 \eta_{ji}^\ell \\ \left(\eta_{ji}^\ell\right)_{\ell=-N}^N \in \text{SOS-II} \end{array} \right\}.
$$

The implemented version of second-order-cone integer program is

$$
\begin{array}{lll}
\max & \sum_{j\in J^+}(\lambda_j-\phi)\sum_{i=1}^r \xi_{ji} - s + r\phi & \\
\text{s.t} & \boldsymbol{V}\in\mathcal{C}',\ (g,\xi,\eta)\in\text{PLA}' & (1) \\
& \sum_{j\in J^-}(\phi-\lambda_j)\sum_{i=1}^r g_{ji}^2 \leq s & (2) \\
& \sum_{i=1}^r g_{ji}^2 \leq \theta_j^2,\ \forall j\in[d] & (3) \\
& \sum_{j\in J^+}(\lambda_j-\phi)\sum_{i=1}^r \xi_{ji} - s + r\phi \leq [\boldsymbol{A}]_{j_1,j_1} + \cdots + [\boldsymbol{A}]_{j_k,j_k} + \sum_{j\in J^+}\frac{r(\lambda_j-\phi)\theta_j^2}{4N^2} & (4) \\
& \sum_{j=1}^d \lambda_j \xi_{ji} \geq \boldsymbol{v}_{i+1}^\top \boldsymbol{A}\boldsymbol{v}_{i+1},\ \forall i\in[r-1] & (5) \\
& \sum_{j=1}^d [\boldsymbol{v}_i]_j \geq 0,\ \forall i\in[r] & (6)
\end{array}
\tag{SOCIP-impl}
$$

where constraint (1) is the convex relaxation and piecewise linear approximation for SPCA combined with thresholding technique that we discussed in Section 2.4 of main content; constraints (2), (3), (4) are the cutting planes; constraints (5), (6) are the symmetric breaking constraints. Here are have

Theorem 5. Given the number of splitting points N, the size of large eigenvalue set $|J^+|$ and its corresponding threshold parameter ϕ, then the SOCIP-impl can be solved within polynomial time.

Proof. The only non-convex part in SOCIP-impl is the SOS-II constraints for variables η. Since the SOS-II constraints imply that: for each $(j,i)\in J^+\times[r]$, at most two continuous variables of $\left(\eta_{ji}^\ell\right)_{\ell=-N}^N$ are non-zero, then there are $2N$ possible situations for non-zero $\left(\eta_{ji}^\ell\right)_{\ell=-N}^N$, and thus $(2N)^{r|J^+|}$ possible situations for all $(j,i)\in J^+\times[r]$. Given a fixed possible situation of SOS-II variables $\left(\eta_{ji}^\ell\right)_{\ell=-N}^N$ for all $(j,i)\in J^+\times[r]$, note that the SOCIP-impl can be transferred into a continuous convex programming (CP) which is able to optimized exactly within polynomial time T (corresponding to the size of input CP, and the additive gap). Therefore the SOCIP-impl can be solved within polynomial time $T \cdot (2N)^{r|J^+|}$ by checking each possible situations. □

Remark 4.1. Note that the Theorem 5 still holds even when the SOCIP-impl only contains constraint (1). The rest of constraints (2)-(6) are used to improve the running time in practice. Thus people can decide which constraint from (2)-(6) is needed when optimizing their own instances.

5. Main Numerical Results

In this section, we present the original numerical results of SOCIP-impl and Baselines on two types of instances described in Section 3 of main content. Three methods for upper bounds are listed as follows:

$$\begin{array}{lll} \text{Baseline-1} := & [\boldsymbol{A}]_{j_1,j_1} + \cdots + [\boldsymbol{A}]_{j_k,j_k}, & \text{where } [\boldsymbol{A}]_{j_1,j_1} \geq [\boldsymbol{A}]_{j_2,j_3} \geq \cdots [\boldsymbol{A}]_{j_d,j_d} \\ \text{Baseline-2} := & \max_{\boldsymbol{P}} \ \text{Tr}\,(\boldsymbol{AP}), & \text{s.t. } \boldsymbol{I}_d \succeq \boldsymbol{P} \succeq \boldsymbol{0},\ \text{Tr}(\boldsymbol{P}) = r,\ \boldsymbol{1}^\top|\boldsymbol{P}|\boldsymbol{1} \leq rk \\ \text{SOCIP-impl} := & \sum_{j\in J^+}(\lambda_j - \phi)\sum_{i=1}^r \xi_{ji} - s + r\phi & \text{s.t. constraint (1), (2), (3), (4), (5)} \end{array}$$

The lower bounds (LB) are computed by feasible solutions of SPCA which obtained from local search algorithm 1 with randomized initialization. The column "Gap" denotes the gap between upper bounds of SOCIP-impl and lower bounds (LB) defined as Gap := $\frac{\text{SOCIP}-\text{LB}}{\text{LB}}$. All original numerical results are reported in the following tables.

Notice that because of the limitation of hardware and software, the Baseline-2 (SDP relaxation method) is hard to scalable since the quadratic increasing of the number of variables in the lifted space. Thus the Baseline-2 (SDP relaxation method) only works for the case Eisen-1 and Eisen-2, and in rest of the tables we remove the column of "Baseline-2".

Para (d, k, r)	LB	Baseline-1	Baseline-2	SOCIP	Gap
$(79, 5, 2)$	15.277	16.295	20.081	16.351	7.028 %
$(79, 10, 2)$	19.705	21.376	21.530	21.184	7.505 %
$(79, 15, 2)$	20.590	24.034	22.036	21.678	5.282%
$(79, 20, 2)$	21.020	25.881	22.197	21.833	3.866%
$(79, 25, 2)$	21.274	27.279	22.203	21.981	3.321%
$(79, 30, 2)$	21.481	28.535	22.203	22.069	2.736%
$(79, 5, 3)$	16.150	16.295	21.998	16.438	1.783 %
$(79, 10, 3)$	20.569	21.376	23.806	21.54	4.721%
$(79, 15, 3)$	21.553	24.034	24.493	24.139	11.997%
$(79, 20, 3)$	21.683	25.881	24.725	24.599	13.447%
$(79, 25, 3)$	23.205	27.279	24.738	24.427	5.268%
$(79, 30, 3)$	23.229	28.535	24.738	24.527	5.588%

Table 1. Compare SOCIP-impl with baseline for Eisen-1

Para (d, k, r)	LB	Baseline-1	Baseline-2	SOCIP	Gap
$(118, 5, 2)$	8.144	8.574	14.008	8.61	5.724%
$(118, 10, 2)$	13.686	15.051	22.211	15.094	10.291%
$(118, 15, 2)$	18.328	20.641	26.984	20.696	12.917%
$(118, 20, 2)$	22.155	25.845	29.322	25.473	14.975%
$(118, 25, 2)$	25.040	30.018	30.786	27.807	11.052%
$(118, 30, 2)$	27.311	33.461	31.814	29.604	8.397%
$(118, 5, 3)$	8.434	8.574	19.328	8.635	2.385%
$(118, 10, 3)$	14.457	15.051	28.708	15.148	4.777%
$(118, 15, 3)$	19.296	20.641	32.086	20.762	7.596%
$(118, 20, 3)$	23.583	25.845	34.152	25.977	10.152%
$(118, 25, 3)$	26.734	30.018	35.545	30.164	12.831%
$(118, 30, 3)$	28.741	33.461	36.495	33.409	16.242%

Table 2. Compare SOCIP-impl with baseline for Eisen-2

Para (d, k, r)	LB	Baseline-1	SOCIP	Gap
$(500, 5, 1)$	1646.454	1720.878	1723.119	4.656%
$(500, 10, 1)$	2641.229	2970.226	2970.658	12.476%
$(500, 20, 1)$	4255.287	5015.718	5007.157	17.669%
$(500, 40, 1)$	6924.530	8280.635	8242.987	19.040%
$(500, 80, 1)$	10741.925	13292.953	13183.996	22.734%
$(500, 120, 1)$	13660.302	17349.272	17165.391	25.659%
$(500, 160, 1)$	15666.335	20599.533	19154.019	22.262%
$(500, 5, 2)$	1709.958	1720.878	1733.222	1.361%
$(500, 10, 2)$	2794.140	2970.226	2989.336	6.986 %
$(500, 20, 2)$	4510.085	5015.718	5045.065	11.862%
$(500, 40, 2)$	7245.277	8280.635	8326.816	14.928%
$(500, 80, 2)$	11226.442	13292.953	13359.151	18.997%
$(500, 120, 2)$	14163.219	17349.272	17429.896	23.065 %
$(500, 160, 2)$	16457.275	20599.533	19070.61	15.880 %

Table 3. Compare SOCIP-impl with baseline for CovColon

Para (d, k, r)	LB	Baseline-1	SOCIP	Gap
$(500, 5, 1)$	4300.497	5177.405	5184.741	20.561%
$(500, 10, 1)$	6008.317	8901.180	8902.889	48.176%
$(500, 20, 1)$	9082.158	15160.617	14641.435	61.211%
$(500, 40, 1)$	13107.045	24293.415	19557.092	49.211 %
$(500, 80, 1)$	17544.277	38358.967	24458.286	39.409%
$(500, 120, 1)$	20797.933	48691.305	27058.292	30.101%
$(500, 160, 1)$	23310.903	57395.584	28527.242	22.377%
$(500, 5, 2)$	4990.132	5177.405	5220.062	4.608%
$(500, 10, 2)$	8125.266	8901.180	8951.928	10.174%
$(500, 20, 2)$	11868.012	15160.617	15226.865	28.302%
$(500, 40, 2)$	16138.886	24293.415	24378.169	51.052%
$(500, 80, 2)$	21396.692	38358.968	30178.957	41.045%
$(500, 120, 2)$	25579.788	48691.305	33116.474	29.463%
$(500, 160, 2)$	28950.488	57395.584	35369.37	22.172%

Table 4. Compare SOCIP-impl with baseline for Lymp

Para (d, k, r)	LB	Baseline-1	SOCIP	Gap
$(1000, 5, 1)$	1665.714	3144.0	1994.758	19.754%
$(1000, 10, 1)$	1766.499	3936.0	2036.647	15.293%
$(1000, 20, 1)$	1834.473	5058.585	2061.048	12.351%
$(1000, 40, 1)$	1918.158	6355.687	2083.412	8.615%
$(1000, 80, 1)$	1993.947	7876.078	2094.868	5.061%
$(1000, 5, 2)$	2208.062	3144.0	2669.481	20.897%
$(1000, 10, 2)$	2358.341	3936.0	2842.534	20.531%
$(1000, 20, 2)$	2583.484	5058.585	2963.481	14.709%
$(1000, 40, 2)$	2762.794	6355.687	3093.255	11.961%
$(1000, 80, 2)$	2681.272	7876.078	3128.75	16.689%
$(1000, 5, 3)$	2650.855	3144.0	6745.96	154.482%
$(1000, 10, 3)$	2839.634	3936.0	7373.92	159.679%
$(1000, 20, 3)$	2996.054	5058.585	7616.909	154.231%
$(1000, 40, 3)$	3361.852	6355.687	7659.558	127.837%
$(1000, 80, 3)$	3657.417	7876.078	7362.175	101.294%

Table 5. Compare SOCIP-impl with baseline for Reddit of size 1000

Para (d, k, r)	LB	Baseline-1	SOCIP	Gap
$(1500, 5, 1)$	1694.560	3327.0	2198.699	29.750%
$(1500, 10, 1)$	1834.871	4344.0	2333.223	27.160%
$(1500, 20, 1)$	1965.926	5649.585	2449.516	24.599%
$(1500, 40, 1)$	2317.149	7316.316	2593.001	11.905%
$(1500, 80, 1)$	2537.644	9129.687	2630.608	3.663%
$(1500, 5, 2)$	2241.333	3327.0	4679.819	108.796%
$(1500, 10, 2)$	2552.943	4344.0	5023.951	96.791%
$(1500, 20, 2)$	3073.350	5649.585	5348.092	74.015%
$(1500, 40, 2)$	3178.420	7316.316	5461.002	71.815%
$(1500, 80, 2)$	3596.531	9129.687	5912.314	64.389%
$(1500, 5, 3)$	2675.918	3327.0	7548.85	182.103%
$(1500, 10, 3)$	3025.420	4344.0	8432.259	178.714%
$(1500, 20, 3)$	3587.461	5649.585	9568.021	166.707%
$(1500, 40, 3)$	3722.302	7316.316	11705.781	214.477%
$(1500, 80, 3)$	4639.675	9129.687	12523.76	169.928%

Table 6. Compare SOCIP-impl with baseline for Reddit of size 1500

Para (d, k, r)	LB	Baseline-1	SOCIP	Gap
$(500, 5, 1)$	55.085	252.808	56.22	2.060 %
$(500, 10, 1)$	56.318	360.936	56.474	0.277%
$(500, 20, 1)$	56.342	371.594	56.474	0.235%
$(500, 40, 1)$	56.370	392.540	56.474	0.184%
$(500, 80, 1)$	56.403	433.838	56.474	0.126%
$(500, 5, 2)$	107.422	252.810	109.825	2.237%
$(500, 10, 2)$	109.971	360.936	110.307	0.306%
$(500, 20, 2)$	110.001	371.594	110.308	0.279%
$(500, 40, 2)$	110.046	392.540	110.309	0.239%
$(500, 80, 2)$	110.109	433.834	113.142	2.754%

Table 7. Compare SOCIP-impl with baseline for block spiked covariance of size 500

Para (d, k, r)	LB	Baseline-1	SOCIP	Gap
$(2000, 5, 1)$	4300.497	5177.405	5184.741	20.561%
$(2000, 10, 1)$	6008.317	8901.180	8902.889	48.176%
$(2000, 20, 1)$	9082.158	15160.617	14641.435	61.211%
$(2000, 40, 1)$	13107.045	24293.415	19557.092	49.211 %
$(2000, 80, 1)$	17544.277	38358.967	24458.286	39.409%
$(2000, 120, 1)$	20797.933	48691.305	27058.292	30.101%
$(2000, 160, 1)$	23310.903	57395.584	28527.242	22.377%
$(2000, 5, 2)$	4990.132	5177.405	5220.062	4.608%
$(2000, 10, 2)$	8125.266	8901.180	8951.928	10.174%
$(2000, 20, 2)$	11868.012	15160.617	15226.865	28.302%
$(2000, 40, 2)$	16138.886	24293.415	24378.169	51.052%
$(2000, 80, 2)$	21396.692	38358.968	30178.957	41.045%
$(2000, 120, 2)$	25579.788	48691.305	33116.474	29.463%
$(2000, 160, 2)$	28950.488	57395.584	35369.37	22.172%

Table 8. Compare SOCIP-impl with baseline for Reddit of size 2000