Appendix

A. Proof of Proposition 1

Proposition 1. *When the cyclic order strategy is used, coordinate descent method is guaranteed to converge to a coordinate-wise minimum of Problem (10) that* $\forall i,\ \mathbf{y}_i^* = \arg\min_{\alpha \geq \hat{L}}\ \mathcal{L}(\mathbf{y}_i^* + \alpha \mathbf{e}_i)$.

Proof. Note that $\mathcal{L}(\mathbf{y})$ is continuous and $\{\mathcal{L}(\mathbf{y}^j)\}$ converges monotonically. Assuming that it converges to $\mathcal{L}^*$ with $\lim_{j\to\infty} \mathcal{L}(\mathbf{y}^j) = \mathcal{L}^*$, we obtain that $\forall \alpha,\ i = 1, .., m$:

$$\mathcal{L}^* = \mathcal{L}(\mathbf{y}^{j-1}) = \mathcal{L}(\mathbf{y}^j) \leq \mathcal{L}(\mathbf{y}^{j-1} + \alpha \mathbf{e}_i). \tag{12}$$

Therefore, the right-handed side in (12) attains it minimum at both 0 and $(\mathbf{y}^j)_i - (\mathbf{y}^{j-1})_i$. Combining with the fact that the subproblem only contains one unique global solution, we have $(\mathbf{y}^{j-1})_i = (\mathbf{y}^j)_i$. Since the coordinate i is picked using cyclic order, we have: $\mathbf{y}^{j-1} = \mathbf{y}^j = \mathbf{y}^*$ and $\mathbf{y}^*$ is a coordinate-wise minimum point. □

B. Proof of Lemma 2

Lemma 2. ***(Sufficient Decrease Condition)*** *It holds that:* $f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) \leq \frac{-\theta \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2}{(\mathbf{x}^{t+1})^T \mathbf{C} \mathbf{x}^{t+1}}$.

Proof. We let B be the working set in the t-th iteration and $N \triangleq \{1, 2, ..., n\} \setminus B$. Since we solve Problem (3) in the t-th iteration, we have:

$$\begin{aligned} & (h(\mathbf{x}_B^{t+1}, \mathbf{x}_N^t) + \tfrac{\theta}{2}\|\mathbf{x}_B^{t+1} - \mathbf{x}_B^t\|_2^2) \ / \ g(\mathbf{x}_B^{t+1}, \mathbf{x}_N^t) \\ \leq\ & (h(\mathbf{z}, \mathbf{x}_N^t) + \tfrac{\theta}{2}\|\mathbf{z} - \mathbf{x}_B^t\|_2^2) \ / \ g(\mathbf{z}, \mathbf{x}_N^t),\ \forall \mathbf{z} \in \mathbb{R}^k. \end{aligned}$$

We let $\mathbf{z} = \mathbf{x}_B^t$ and combine with the fact that $\mathbf{x}_N^{t+1} = \mathbf{x}_N^t$, we have:

$$\begin{aligned} & (h(\mathbf{x}_B^{t+1}, \mathbf{x}_N^{t+1}) + \tfrac{\theta}{2}\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2) \ / \ g(\mathbf{x}_B^{t+1}, \mathbf{x}_N^{t+1}) \\ \leq\ & (h(\mathbf{x}_B^t, \mathbf{x}_N^t) + 0) \ / \ g(\mathbf{x}_B^t, \mathbf{x}_N^t). \end{aligned}$$

Noticing the fact that $h(\mathbf{x}_B^t, \mathbf{x}_N^t) = \frac{1}{2}(\mathbf{x}^t)^T \mathbf{A} \mathbf{x}^t$ and $g(\mathbf{x}_B^t, \mathbf{x}_N^t) = \frac{1}{2}(\mathbf{x}^t)^T \mathbf{C} \mathbf{x}^t$, we have:

$$\begin{aligned} & ((\mathbf{x}^{t+1})^T \mathbf{A} \mathbf{x}^{t+1} + \theta \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2) \ / \ ((\mathbf{x}^{t+1})^T \mathbf{C} \mathbf{x}^{t+1}) \\ \leq\ & ((\mathbf{x}^t)^T \mathbf{A} \mathbf{x}^t) \ / \ ((\mathbf{x}^t)^T \mathbf{C} \mathbf{x}^t). \end{aligned}$$

Moreover, using the structure of the objective function $f(\cdot)$, we obtain: $f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) = \frac{(\mathbf{x}^{t+1})^T \mathbf{A} \mathbf{x}^{t+1}}{(\mathbf{x}^{t+1})^T \mathbf{C} \mathbf{x}^{t+1}} - \frac{(\mathbf{x}^t)^T \mathbf{A} \mathbf{x}^t}{(\mathbf{x}^t)^T \mathbf{C} \mathbf{x}^t} \leq \frac{-\theta \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2}{(\mathbf{x}^{t+1})^T \mathbf{C} \mathbf{x}^{t+1}}$. Thus, we finish the proof of this lemma.

□

C. Proof of Theorem 2

We now prove the convergence properties of Algorithm 1. The following supermartingale convergence result is useful in our analysis [31].

Lemma 3. *[31] Let* $\mathbf{v}_t$, $\mathbf{u}_t$ *and* $\boldsymbol{\alpha}_t$ *be three sequences of nonnegative random variables such that*

$$\begin{aligned} & \mathbb{E}[\mathbf{v}_{t+1} \mid \mathcal{F}_t] \leq (1 + \boldsymbol{\alpha}_t)\mathbf{v}_t - \mathbf{u}_t,\ \forall t \geq 0\ a.s. \\ & \text{and } \textstyle\sum_{t=0}^{\infty}\ \boldsymbol{\alpha}_t < \infty\ a.s., \end{aligned} \tag{13}$$

where $\mathcal{F}_t$ *denotes the collections* $\{\mathbf{v}_0, ..., \mathbf{v}_t,\ \mathbf{u}_0, ..., \mathbf{u}_t, \boldsymbol{\alpha}_0, ..., \boldsymbol{\alpha}_t\}$. *Then, we have* $\lim_{t\to\infty} \mathbf{v}_t = \chi$ *for a random variable* $\chi \geq 0$ *a.s. and* $\sum_{t=0}^{\infty} \mathbf{u}_t < \infty$ *a.s.*

We now present our main results.

Theorem 2. ***Convergence Properties of Algorithm 1.*** *Assume that the subproblem in (3) is solved globally, and there exists a constant* σ *such that* $\mathbf{x}^t \mathbf{C} \mathbf{x}^t \geq \sigma > 0$ *for all* t. *We have the following results.*

(i) *When the random strategy is used to find the working set, we have* $\lim_{t\to\infty} \mathbb{E}[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|] = 0$ *and Algorithm 1 converges to the block-*k *stationary point in expectation.*

(ii) *When the swapping strategy is used to find the working set with* $k \geq 2$, *we have* $\lim_{t\to\infty} \|\mathbf{x}^{t+1} - \mathbf{x}^t\| = 0$ *and Algorithm 1 converges to the block-2 stationary point deterministically.*

Proof. We use $\mathbf{x}^*$ and $\bar{\mathbf{x}}$ to denote any optimal point and any block-k stationary point of (1), respectively. We use the notation ξ^t for the entire history of random index selection:

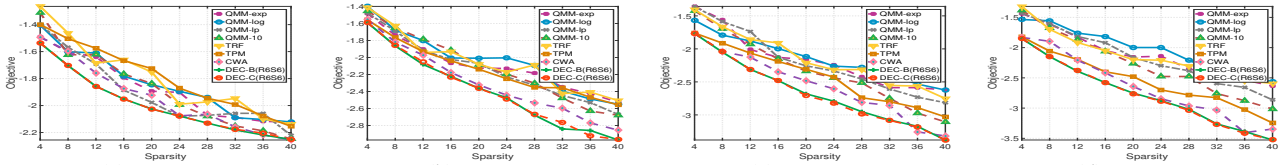
$$\xi^t = \{B^0,\ B^1,\ ..., B^t\}$$

(i) We notice that B^t is independent on the past B^{t-1}, while $\mathbf{x}^t$ fully depends on ξ^{t-1}. Taking the expectation conditioned on ξ^{t-1} for the sufficient descent inequality in Lemma 2, we obtain:

$$\begin{aligned} & \mathbb{E}[f(\mathbf{x}^{t+1}) | \xi^t] - f(\mathbf{x}^t) \\ \leq\ & -\mathbb{E}[\tfrac{\theta \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2}{(\mathbf{x}^{t+1})^T \mathbf{C} \mathbf{x}^{t+1}} | \xi^t] \\ \overset{(a)}{\leq}\ & -\tfrac{\theta}{\sigma} \mathbb{E}[\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 | \xi^t] \\ =\ & -\tfrac{\theta}{\sigma} \tfrac{1}{C_n^k} \textstyle\sum_{i=1}^{C_n^k}\ \|\mathcal{P}(\mathcal{B}_{(i)},\ \mathbf{x}^t) - \mathbf{x}^t_{\mathcal{B}_{(i)}}\|_2^2 \\ \overset{(b)}{=}\ & -\tfrac{\theta}{\sigma} \cdot \mathcal{M}(\mathbf{x}^t) \end{aligned} \tag{14}$$

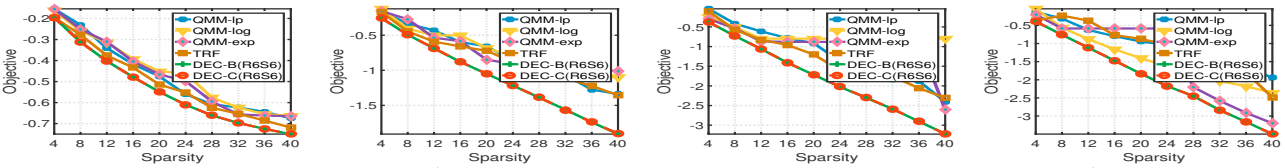
step (a) uses the assumption that $\mathbf{x}^t \mathbf{C} \mathbf{x}^t \geq \sigma > 0,\ \forall \mathbf{x}^t$ which clearly holds since $\mathbf{C}$ is strictly positive and $\mathbf{x}^t \neq \mathbf{0}$; step (b) uses the definition of $\mathcal{M}(\mathbf{x}^t)$ in Definition 1. Therefore, we have:

$$\mathbb{E}[f(\mathbf{x}^{t+1}) \mid \xi^t] - f(\mathbf{x}^*) \leq f(\mathbf{x}^t) - f(\mathbf{x}^*) - \tfrac{\theta \mathcal{M}(\mathbf{x}^t)}{\sigma} \tag{15}$$



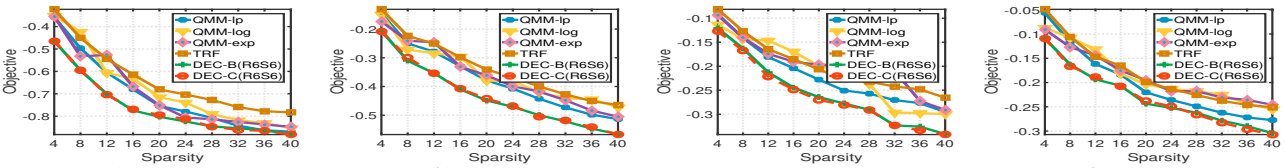
(a) randn-100 (b) randn-500 (c) randn-1500 (d) randn-2000

Figure 6 Accuracy of different methods on different data sets for sparse PCA problem with varying the cardinalities.



(a) randn-100 (b) randn-500 (c) randn-1500 (d) randn-2000

Figure 7 Accuracy of different methods on different data sets for sparse FDA problem with varying the cardinalities.



(a) randn-100 (b) randn-500 (c) randn-1500 (d) randn-2000

Figure 8 Accuracy of different methods on different data sets for sparse CCA problem with varying the cardinalities.

Using the supermartingale convergence theorem given in Lemma 3 with $\mathbf{v}_t = \mathbb{E}[f(\mathbf{x}^{t+1}) \mid \xi^t] - f(\mathbf{x}^*) \geq 0$ and $\mathbf{u}_t = \frac{\theta \mathcal{M}(\mathbf{x}^t)}{\sigma}$, we have

$$\lim_{t\to\infty} f(\mathbf{x}^t) - f(\mathbf{x}^*) = \chi \ a.s.$$

for a certain random variable $\chi \geq 0$ and thus the sequence $f(\mathbf{x}^t)$ converges to a random variable $\bar{F} = \chi + f(\mathbf{x}^*)$. In addition, we have $\lim_{t\to\infty} f(\mathbf{x}^t) - f(\mathbf{x}^{t+1}) = 0$ almost surely. From (14), we have

$$\lim_{t\to\infty} \mathcal{M}(\mathbf{x}^t) = 0, \quad \lim_{t\to\infty} \|\mathbf{x}^t - \mathbf{x}^{t+1}\| = 0.$$

Therefore, the algorithm converges to the block-k stationary point. Summing the inequality in (14) over $i = 0, 1, ..., t-1$, we have:

$$\frac{\theta}{\sigma} \cdot \sum_{i=0}^{t} \mathcal{M}(\mathbf{x}^i) \leq f(\mathbf{x}^0) - f(\mathbf{x}^t).$$

Using the fact that $f(\mathbf{x}^*) \leq f(\mathbf{x}^t)$, we obtain:

$$\begin{aligned} & \frac{\theta}{\sigma} \sum_{i=0}^{t} \mathbb{E}[\|\mathcal{M}(\mathbf{x}^i) \mid \xi^i] \leq f(\mathbf{x}^0) - f(\mathbf{x}^*) \\ \Rightarrow \quad & \min_{i=1,\ldots,t} \mathbb{E}[\mathcal{M}(\mathbf{x}^i) \mid \xi^i] \leq \frac{\sigma(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{t\theta}. \end{aligned}$$

We conclude that $\mathbf{x}^t$ converges to the block-k stationary point with $\min_{i=1,\ldots,t} \mathbb{E}[\mathcal{M}(\mathbf{x}^i) \mid \mathbf{x}^i] \leq \mathcal{O}(1/t)$.

(ii) We now prove the second part of this theorem. We have the following inequalities:

$$\begin{aligned} f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) &\leq -\frac{\theta \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2}{(\mathbf{x}^{t+1})^T \mathbf{C} \mathbf{x}^{t+1}} \\ &\leq -\frac{\theta}{\sigma} \|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2^2 \end{aligned}$$

Summing this inequality over $i = 0, 1, ..., t-1$, we have:

$$\begin{aligned} & \frac{\theta}{\sigma} \cdot \sum_{i=0}^{t} \|\mathbf{x}^{i+1} - \mathbf{x}^i\|_2^2 \leq f(\mathbf{x}^0) - f(\mathbf{x}^t) \\ \Rightarrow \quad & \min_{i=1,\ldots,t} \|\mathbf{x}^{i+1} - \mathbf{x}^i\|_2^2 \leq \frac{\sigma}{\theta} \frac{f(\mathbf{x}^0) - f(\mathbf{x}^*)}{t}. \end{aligned}$$

Using the fact that $f(\mathbf{x}^*) \leq f(\mathbf{x}^t)$, we have $\lim_{t\to\infty} \|\mathbf{x}^{t+1} - \mathbf{x}^t\| = 0$. Therefore, Algorithm 1 is convergent when swapping strategy is used.

We now prove that Algorithm 1 convergence to a block-2 stationary point $\bar{\mathbf{x}}$. Since Algorithm 1 is monotonically non-increasing and converges to a stationary point $\bar{\mathbf{x}}$ such that no decrease is made, we have $\mathbf{D}_{i,j} \geq 0$ for (4). Therefore, it holds that $\min_\alpha f(\bar{\mathbf{x}} + \alpha \mathbf{e}_i - (\bar{\mathbf{x}})_j \mathbf{e}_j) \geq f(\bar{\mathbf{x}})$, $\forall i \in \bar{\mathcal{S}}(\bar{\mathbf{x}})$, $j \in \bar{\mathcal{Z}}(\bar{\mathbf{x}})$. We have the following result: $f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}} + \mathbf{d}), \forall \mathbf{d}$ with $\|\mathbf{d} - \bar{\mathbf{x}}\|_0 = 2$. Therefore, $\bar{\mathbf{x}}$ is a block-2 stationary point. □

D. Additional Experiments

We demonstrate the experimental results on the randomized generated data sets for sparse PCA, sparse FDA, and sparse CCA in Figure 6, 7 and 8, respectively. These results further consolidate our conclusions drawn in Section 7.