Supplementary Material for Unsupervised Discriminative Feature Selection With $\ell_{2,0}$-Norm Constrained Sparse Projection

Xia Dong, Feiping Nie*, *Senior Member, IEEE*, Lai Tian, Rong Wang, and Xuelong Li, *Fellow, IEEE*

I. Notations

TABLE I: Summary of Notations

Notations	Descriptions
n	Number of samples
d	Number of features
c	Number of clusters
m	Reduced dimensionality
k	Number of selected features
$\mathbf{1}_n$	Vector with all n elements as one
$\boldsymbol{I}_{n\times n}$	Identity matrix with size $n \times n$
$\mathbb{R}$	Set of real numbers
$\mathbb{Z}^+$	Set of positive integers
$\mathrm{Tr}(\boldsymbol{X})$	Trace of square matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$
$\mathrm{rank}(\boldsymbol{X})$	Rank of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\boldsymbol{x}_i$	The i-th column of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
x_{ij}	The (ij)-th element of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\boldsymbol{X}^\top$	Transpose of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\boldsymbol{X}^\dagger$	Moore-Penrose inverse of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\{\lambda_i(\boldsymbol{X})\}_{i=1}^n$	Eigenvalues of $\mathbf{X}$, ordered in descending order
$\|\boldsymbol{X}\|_F = \sqrt{\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X})}$	Frobenius norm of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\|\boldsymbol{X}\|_{p,q} = \left(\sum_{i=1}^n \|\boldsymbol{x}^i\|_p^q\right)^{1/q}$	$\ell_{p,q}$-norm of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\|\boldsymbol{X}\|_1 = \max_{j\in[1,d]} \sum_{i=1}^n |x_{ij}|$	1-norm of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$
$\|\boldsymbol{X}\|_\infty = \max_{i\in[1,n]} \sum_{j=1}^d |x_{ij}|$	Infinity norm of matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$

II. An Example of Matrix A

To clarify the description of matrix $\boldsymbol{A} \in \{0,1\}^{d\times k}$ in Section 4.1.1, we provide an example. Suppose there are $d = 6$ inputs, and we select $k = 3$ with row indices $\boldsymbol{q} = [2, 4, 5]$. According to the definition of the operator $\Omega_d^k(\boldsymbol{q})$, the corresponding row-selection matrix $\boldsymbol{A}$ is:

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

From this example, we see that $\boldsymbol{A}$ is a sparse matrix with k columns, each containing exactly one 1 at the row index specified by $\boldsymbol{q}$, which implies that $\boldsymbol{A}^\top \mathbf{1}_d = \mathbf{1}_k$.

III. The Proof of Theorem 1

Proof. Suppose $\boldsymbol{x}_*$ is the globally optimal solution to problem (7), with the corresponding globally minimal objective value α_*. This implies that $\frac{h(\boldsymbol{x}_*)}{p(\boldsymbol{x}_*)} = \alpha_*$. Consequently, $\forall\ \boldsymbol{x} \in \mathcal{C}$, we have $\frac{h(\boldsymbol{x})}{p(\boldsymbol{x})} \geq \alpha_*$. Since $p(\boldsymbol{x}) > 0$, it follows that $h(\boldsymbol{x}) - \alpha_* p(\boldsymbol{x}) \geq 0$. Moreover, noting that $h(\boldsymbol{x}_*) - \alpha_* p(\boldsymbol{x}_*) = 0$, we conclude that $\min_{\boldsymbol{x}\in\mathcal{C}} \left(h(\boldsymbol{x}) - \alpha_* p(\boldsymbol{x})\right) = 0$. Now, define the function $f(\alpha) = \min_{\boldsymbol{x}\in\mathcal{C}} \left(h(\boldsymbol{x}) - \alpha p(\boldsymbol{x})\right)$. Then, we have $f(\alpha_*) = 0$. This completes the proof. □

IV. The Proof of Theorem 2

Proof. In Algorithm 1, we observe from lines 1–2 that $h(\boldsymbol{x}_t) - \alpha_t p(\boldsymbol{x}_t) = 0$ and $h(\boldsymbol{x}_{t+1}) - \alpha_t p(\boldsymbol{x}_{t+1}) \leq h(\boldsymbol{x}_t) - \alpha_t p(\boldsymbol{x}_t)$. Accordingly, it follows that $h(\boldsymbol{x}_{t+1}) - \alpha_t p(\boldsymbol{x}_{t+1}) \leq 0$, which implies $\frac{h(\boldsymbol{x}_{t+1})}{p(\boldsymbol{x}_{t+1})} \leq \alpha_t = \frac{h(\boldsymbol{x}_t)}{p(\boldsymbol{x}_t)}$. This indicates that Algorithm 1 guarantees the objective function of problem (7) is non-increasing at each iteration until convergence.

According to Theorem 1, the global minimum of the objective in problem (7) corresponds to the root of the function $f(\alpha)$. It is well known that Newton's method is widely regarded as an effective algorithm for root-finding under standard regularity conditions. According to line 2 of Algorithm 1, let $f(\alpha_t) = h(\boldsymbol{x}_{t+1}) - \alpha_t p(\boldsymbol{x}_{t+1})$, then the derivative is $f'(\alpha_t) = -p(\boldsymbol{x}_{t+1})$. Applying the Newton's update rule, we obtain

$$\alpha_{t+1} = \alpha_t - \frac{f(\alpha_t)}{f'(\alpha_t)} = \alpha_t - \frac{h(\boldsymbol{x}_{t+1}) - \alpha_t p(\boldsymbol{x}_{t+1})}{-p(\boldsymbol{x}_{t+1})} = \frac{h(\boldsymbol{x}_{t+1})}{p(\boldsymbol{x}_{t+1})},$$

which coincides with line 1 of Algorithm 1. Therefore, the iterative scheme in Algorithm 1 is equivalent to applying Newton's method to find the root of $f(\alpha)$. According to [1], Newton's method enjoys a quadratic convergence rate under standard regularity conditions. This completes the proof. □

V. The Proof of Remark 1

Proof. According to problems (16) and (22), we have

$$f(\boldsymbol{W}_t) = \mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right), \tag{1}$$

$$g\left(\boldsymbol{W}_t|\boldsymbol{W}_t\right) = \mathrm{Tr}\left(\boldsymbol{W}_t^\top \left(\boldsymbol{S}_d \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{S}_d\right) \boldsymbol{W}_t\right). \tag{2}$$

It is straightforward to verify that $f(\boldsymbol{W}_t) = g\left(\boldsymbol{W}_t|\boldsymbol{W}_t\right)$ since $\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^\dagger\boldsymbol{P}$ for any matrix $\boldsymbol{P}$.

Since $\boldsymbol{S}_d$ is positive semi-definite, it admits a factorization $\boldsymbol{S}_d = \boldsymbol{Q}\boldsymbol{Q}^\top$. Denote the following matrices:

$$\boldsymbol{\Upsilon} = \boldsymbol{Q}^\top \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{Q}, \tag{3}$$
$$\boldsymbol{\Psi} = \boldsymbol{Q}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{Q}. \tag{4}$$

Then, the function $g\left(\boldsymbol{W}_t|\boldsymbol{W}_t\right)$ can be rewritten as

$$\begin{aligned} g\left(\boldsymbol{W}|\boldsymbol{W}_t\right) &= \operatorname{Tr}\left(\boldsymbol{W}^\top \left(\boldsymbol{S}_d \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{S}_d\right) \boldsymbol{W}\right) \\ &= \operatorname{Tr}\left(\boldsymbol{\Upsilon}\boldsymbol{\Psi}\right). \end{aligned} \tag{5}$$

According to Theorems 4.3.53 and 1.3.22 [2], and noting that $\lambda_i\left(\boldsymbol{\Psi}\right) \geq 0$ for all $i \in [1, m]$, we obtain

$$\operatorname{Tr}\left(\boldsymbol{\Upsilon}\boldsymbol{\Psi}\right) \leq \sum_{i=1}^{d} \lambda_i\left(\boldsymbol{\Upsilon}\right) \lambda_i\left(\boldsymbol{\Psi}\right) \leq \sum_{i=1}^{m} \lambda_i\left(\boldsymbol{\Psi}\right). \tag{6}$$

Since $\operatorname{rank}\left(\boldsymbol{\Psi}\right) \leq \operatorname{rank}\left(\boldsymbol{W}\right) = m$, we have $\sum_{i=1}^{m} \lambda_i\left(\boldsymbol{\Psi}\right) = \operatorname{Tr}\left(\boldsymbol{\Psi}\right)$. That is, $\operatorname{Tr}\left(\boldsymbol{\Upsilon}\boldsymbol{\Psi}\right) \leq \operatorname{Tr}\left(\boldsymbol{\Psi}\right) = \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d \boldsymbol{W}\right) = f(\boldsymbol{W})$. In summary, we have $g\left(\boldsymbol{W}|\boldsymbol{W}_t\right) \leq f(\boldsymbol{W})$. This completes the proof. □

VI. The Proof of Theorem 3

According to [3], we provide the proof of Theorem 3 below.

Proof. Recall that Remark 1 demonstrates that the surrogate problem (22) for optimizing $\boldsymbol{W}$ meets the condition (20) required by the majorize-minimization (MM) framework [4], [5]. Let $\widetilde{\boldsymbol{W}}_{t+1} = \arg\max_{\boldsymbol{W}} g(\boldsymbol{W}|\boldsymbol{W}_t)$, according to Eq. (21), the following inequality holds:

$$f(\widetilde{\boldsymbol{W}}_{t+1}) \geq g(\widetilde{\boldsymbol{W}}_{t+1}|\boldsymbol{W}_t) \geq g(\boldsymbol{W}_t|\boldsymbol{W}_t) = f(\boldsymbol{W}_t). \tag{7}$$

According to Eq. (1) and inequality (7), we have

$$\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \leq \operatorname{Tr}\left(\widetilde{\boldsymbol{W}}_{t+1}^\top \boldsymbol{S}_d \widetilde{\boldsymbol{W}}_{t+1}\right). \tag{8}$$

Given $\widetilde{\boldsymbol{W}}_{t+1} = \boldsymbol{A}_{t+1}\widetilde{\boldsymbol{B}}_{t+1}$ and $\boldsymbol{W}_{t+1} = \boldsymbol{A}_{t+1}\boldsymbol{B}_{t+1}$. According to problem (23), $\boldsymbol{B}_{t+1}$ maximizes its objective in the $(t+1)$-th iteration, then we have

$$\begin{aligned} \operatorname{Tr}\left(\widetilde{\boldsymbol{W}}_{t+1}^\top \boldsymbol{S}_d \widetilde{\boldsymbol{W}}_{t+1}\right) &= \operatorname{Tr}\left(\widetilde{\boldsymbol{B}}_{t+1}^\top \boldsymbol{A}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{A}_{t+1} \widetilde{\boldsymbol{B}}_{t+1}\right) \\ &\leq \operatorname{Tr}\left(\boldsymbol{B}_{t+1}^\top \boldsymbol{A}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{A}_{t+1} \boldsymbol{B}_{t+1}\right) \\ &= \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right). \end{aligned} \tag{9}$$

According to inequalities (8) and (9), we have

$$\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \leq \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right). \tag{10}$$

This indicates that Algorithm 2 ensures the objective of problem (16) remains non-decreasing with each iteration. Then we aim to prove that if $\boldsymbol{W}_t \neq \boldsymbol{W}_{t+1}$, then $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \neq \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$. This result demonstrates the ascent property of Algorithm 2, namely, $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) < \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$.

Note that if $\boldsymbol{W}_t \neq \boldsymbol{W}_{t+1}$, then $\boldsymbol{A}_t \neq \boldsymbol{A}_{t+1}$, since $\boldsymbol{W} = \boldsymbol{A}\boldsymbol{B}$ and $\boldsymbol{B}$ is formed by the leading m eigenvectors of $\left(\boldsymbol{A}^\top \boldsymbol{S}_d \boldsymbol{A}\right)$. Therefore, suppose that there exists $\boldsymbol{A}_t \neq \boldsymbol{A}_{t+1}$ such that $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$. Then the equality in inequality (6) holds. According to the equality condition in Theorem 4.3.53 [2], the matrices $\boldsymbol{Q}^\top \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{Q}$ and $\boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{W}_{t+1}^\top \boldsymbol{Q}$ are simultaneously diagonalizable. Assuming that $\boldsymbol{S}_d$ is full rank, we have that $\boldsymbol{\Omega}_t = \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t$ is diagonal. Define $\boldsymbol{\Phi}_t = \boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2}$, then

$$\boldsymbol{Q}^\top \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1} \boldsymbol{W}_t^\top \boldsymbol{Q} = \boldsymbol{\Phi}_t \boldsymbol{\Phi}_t^\top, \tag{11}$$

$$\boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t = \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\Omega}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{I}_{m\times m}. \tag{12}$$

From the simultaneously diagonalizable property and Theorem 1.3.22 [2], it follows that

$$\begin{aligned} \boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{W}_{t+1}^\top \boldsymbol{Q} &= \boldsymbol{\Phi}_t \boldsymbol{\Omega}_{t+1} \boldsymbol{\Phi}_t^\top \\ &= \boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\Omega}_{t+1} \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{W}_t^\top \boldsymbol{Q}. \end{aligned} \tag{13}$$

Based on Eq. (13), we have

$$\begin{aligned} & \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t \\ =\ & \boldsymbol{W}_t^\top \boldsymbol{Q} \left(\boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{W}_{t+1}^\top \boldsymbol{Q}\right) \boldsymbol{Q}^\top \boldsymbol{W}_t \\ =\ & \boldsymbol{W}_t^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\Omega}_{t+1} \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{W}_t^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_t \\ =\ & \boldsymbol{\Omega}_t \boldsymbol{\Omega}_{t+1}. \end{aligned} \tag{14}$$

We now consider the objective of the surrogate problem (22), leading to

$$\begin{aligned} & \operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)^\dagger \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ =\ & \operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ =\ & \operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1} \boldsymbol{W}_{t+1}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_t\right) \\ =\ & \operatorname{Tr}\left(\boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1} \boldsymbol{W}_{t+1}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{Q}\right) \\ =\ & \operatorname{Tr}\left(\boldsymbol{\Upsilon}_{t+1} \boldsymbol{\Psi}_t\right). \end{aligned} \tag{15}$$

From Eq. (15), inequality (6), and Theorem 4.3.53 [2], we obtain

$$\begin{aligned} & \operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ \leq\ & \operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \operatorname{Tr}\left(\boldsymbol{\Omega}_t\right). \end{aligned} \tag{16}$$

Let $\boldsymbol{\Gamma} = \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\mho} = \boldsymbol{\Omega}_{t+1}^{-1} \in \mathbb{R}^{m\times m}$, then based on Theorem 4.3.53 [2], we have

$$\begin{aligned} & \operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ =\ & \operatorname{Tr}\left(\boldsymbol{\Gamma}\boldsymbol{\mho}\right) \geq \sum_{i=1}^{m} \lambda_i(\boldsymbol{\Gamma}) \lambda_{m-i+1}(\boldsymbol{\mho}). \end{aligned} \tag{17}$$

Note that $\lambda_{m-i+1}(\boldsymbol{\mho}) = \lambda_i(\boldsymbol{\Omega}_{t+1})^{-1}$, and by Eq. (14) and Theorem 1.3.22 [2], we get

$$\sum_{i=1}^{m} \lambda_i(\boldsymbol{\Gamma}) \lambda_{m-i+1}(\boldsymbol{\mho}) = \sum_{i=1}^{m} \frac{\lambda_i(\boldsymbol{\Omega}_t \boldsymbol{\Omega}_{t+1})}{\lambda_i(\boldsymbol{\Omega}_{t+1})} = \operatorname{Tr}\left(\boldsymbol{\Omega}_t\right). \tag{18}$$

Combing inequalities (16), (17) and Eq. (18), we conclude that $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \operatorname{Tr}\left(\boldsymbol{\Omega}_t\right)$. According to Theorem 4.3.53 [2], the equality in inequality (17) implies that $\boldsymbol{\Gamma}$ and $\boldsymbol{\mho}$ are simultaneously diagonalizable. From Eq. (14), we know that $\boldsymbol{\Gamma} = \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} = \boldsymbol{\Omega}_t \boldsymbol{\Omega}_{t+1}$, which implies that $\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t = \boldsymbol{\Omega}_t^{1/2} \boldsymbol{\Omega}_{t+1}^{1/2}$ is diagonal. Recall that

$\boldsymbol{\Phi}_t = \boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2}$ and $\boldsymbol{\Phi}_{t+1} = \boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1/2}$, then we have

$$\begin{aligned}
&\boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t = \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{I}_{m\times m}, \\
&\boldsymbol{\Phi}_{t+1}^\top \boldsymbol{\Phi}_{t+1} = \boldsymbol{\Omega}_{t+1}^{-1/2} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1/2} = \boldsymbol{I}_{m\times m}, \\
&\boldsymbol{\Phi}_{t+1}^\top \boldsymbol{\Phi}_t = \boldsymbol{\Omega}_{t+1}^{-1/2} \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{I}_{m\times m}.
\end{aligned} \tag{19}$$

Thus, we conclude that $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}_{t+1}$, which implies that $\boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{S}_d \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1/2}$. Since $\boldsymbol{S}_d$ is full rank, it follows that $\boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{W}_{t+1} \boldsymbol{\Omega}_{t+1}^{-1/2}$. Therefore, we conclude that $\boldsymbol{A}_t = \boldsymbol{A}_{t+1}$, since the operation $\boldsymbol{\Omega}^{-1/2}$ does not affect the sparsity pattern of $\boldsymbol{W}$. This leads to a contradiction with our initial assumption.

The above analysis establishes the non-decreasing property of the sequence $\{\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)\}_{t\in\mathcal{Z}^+}$. Note that $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)$ is upper bounded by $\mathrm{Tr}\left(\boldsymbol{S}_d\right)$. Therefore, the sequence will eventually converge after a finite number of iterations. Moreover, we have shown that if $\boldsymbol{W}_t \neq \boldsymbol{W}_{t+1}$, then $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) < \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$. Therefore, by the contrapositive, if the objective value converges, i.e., $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$, then it must be that $\boldsymbol{W}_t = \boldsymbol{W}_{t+1}$ and $\boldsymbol{A}_t = \boldsymbol{A}_{t+1}$. Consequently, the sequence $\{\boldsymbol{W}_t\}_{t\in\mathcal{Z}^+}$ converges to a fixed point $\widehat{\boldsymbol{W}}$, and as $\boldsymbol{W}_t \to \widehat{\boldsymbol{W}}$, we have $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \to \mathrm{Tr}\left(\widehat{\boldsymbol{W}}^\top \boldsymbol{S}_d \widehat{\boldsymbol{W}}\right)$. This completes the proof. □

VII. The Proof of Theorem 4

Proof. In Algorithm 3, the optimization strategy for problem (6) consists of an outer loop and a two-layer inner loop. The two-layer inner loop involves three optimization variables: $\boldsymbol{W}$, $\boldsymbol{M}$, and $\boldsymbol{Y}$. For $\boldsymbol{M}$ and $\boldsymbol{Y}$, we can get their closed-form solutions through Eqs. (12) and (26), respectively. For $\boldsymbol{W}$, when $\mathrm{rank}\left(\boldsymbol{S}_d\right) \leq m$, we can get the globally optimal solution through Algorithm 2; When $\mathrm{rank}\left(\boldsymbol{S}_d\right) > m$, Theorem 3 guarantees that Algorithm 2 achieves convergence in both the objective and the iterates. For the outer loop, Theorem 2 shows that the quadratic convergence of the objective in the ratio minimization problem (7) ensures the convergence of the objective in problem (6). Specifically, assume that the optimization variables in the t-th iteration are $\boldsymbol{W}_t$, $\boldsymbol{M}_t$, and $\boldsymbol{Y}_t$, and in the $(t+1)$-th iteration are $\boldsymbol{W}_{t+1}$, $\boldsymbol{M}_{t+1}$, and $\boldsymbol{Y}_{t+1}$. Let $\mathcal{L}(\boldsymbol{W}, \boldsymbol{M}, \boldsymbol{Y}) = \sum_{i=1}^n \sum_{j=1}^c y_{ij}^r \|\boldsymbol{W}^\top \boldsymbol{x}_i - \boldsymbol{m}_j\|_2^2$ and $\mathcal{L}(\boldsymbol{W}) = \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_t \boldsymbol{W}\right)$. Then, we have

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{W}_{t+1}, \boldsymbol{M}_{t+1}, \boldsymbol{Y}_{t+1}\right) - \alpha_t \mathcal{L}\left(\boldsymbol{W}_{t+1}\right) \\
\leq\ &\mathcal{L}\left(\boldsymbol{W}_t, \boldsymbol{M}_t, \boldsymbol{Y}_t\right) - \alpha_t \mathcal{L}\left(\boldsymbol{W}_t\right) = 0,
\end{aligned}$$

and

$$\frac{\mathcal{L}\left(\boldsymbol{W}_{t+1}, \boldsymbol{M}_{t+1}, \boldsymbol{Y}_{t+1}\right)}{\mathcal{L}\left(\boldsymbol{W}_{t+1}\right)} \leq \alpha_t = \frac{\mathcal{L}\left(\boldsymbol{W}_t, \boldsymbol{M}_t, \boldsymbol{Y}_t\right)}{\mathcal{L}\left(\boldsymbol{W}_t\right)}.$$

This indicates that Algorithm 3 ensures the objective of problem (6) remains non-increasing with each iteration and ultimately converges at the objective level. This completes the proof. □

VIII. The Proof of Lemma 1

According to [3], we provide the proof of Lemma 1 below.

Proof. Let $\boldsymbol{S}_d = \boldsymbol{S}_d^m + \boldsymbol{S}_d^c$, We then have the following derivation:

$$\begin{aligned}
&\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) \\
=&\max_{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m},\ \|\boldsymbol{W}\|_{2,0}=k} \mathrm{Tr}\left(\boldsymbol{W}^\top \left(\boldsymbol{S}_d^m + \boldsymbol{S}_d^c\right) \boldsymbol{W}\right) \\
\leq&\max_{\substack{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m},\\ \|\boldsymbol{W}\|_{2,0}=k}} \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^m \boldsymbol{W}\right) + \max_{\substack{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m},\\ \|\boldsymbol{W}\|_{2,0}=k}} \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^c \boldsymbol{W}\right) \\
\leq&\ \mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) + \max_{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m}} \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^c \boldsymbol{W}\right) \\
\leq&\ \mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) + \sum_{i=m+1}^{2m} \lambda_i(\boldsymbol{S}_d),
\end{aligned} \tag{20}$$

from which it follows that

$$\frac{\mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right)}{\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right)} \geq 1 - \frac{\sum_{i=m+1}^{2m} \lambda_i(\boldsymbol{S}_d)}{\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right)}. \tag{21}$$

Furthermore, we have the following two observations:

$$\begin{aligned}
\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) &\overset{(a)}{\geq} \max_{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m},\ \|\boldsymbol{W}\|_{2,0}=m} \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d \boldsymbol{W}\right) \\
&\overset{(b)}{\geq} \frac{m}{d} \mathrm{Tr}\left(\boldsymbol{S}_d\right) = \frac{m}{d} \sum_{i=1}^d \lambda_i(\boldsymbol{S}_d),
\end{aligned} \tag{22}$$

$$\begin{aligned}
\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) &\overset{(c)}{\geq} \max_{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m},\ \|\boldsymbol{W}\|_{2,0}=k} \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^m \boldsymbol{W}\right) \\
&\overset{(d)}{\geq} \frac{k}{d} \sum_{i=1}^d \lambda_i(\boldsymbol{S}_d^m) = \frac{k}{d} \sum_{i=1}^m \lambda_i(\boldsymbol{S}_d),
\end{aligned} \tag{23}$$

where (a) holds because $m \leq k$, and (c) holds since $\boldsymbol{S}_d$ is positive semi-definite. Inequality (b) holds since $k = m$, implying that the problem on the left side of (b) achieves its global optimum via Algorithm 2. Inequality (d) holds because $\mathrm{rank}(\boldsymbol{S}_d^m) = m$, and thus the problem on the left side of (d) also achieves its global optimum via Algorithm 2.

Let $\mathrm{z} = \min\{\mathrm{rank}\left(\boldsymbol{S}_d\right), 2m\}$, $\mathrm{c}_1 = \frac{\sum_{i=m+1}^{\mathrm{z}} \lambda_i(\boldsymbol{S}_d)}{\sum_{i=1}^m \lambda_i(\boldsymbol{S}_d)}$, and $\mathrm{c}_2 = \frac{\sum_{i=m+1}^{\mathrm{z}} \lambda_i(\boldsymbol{S}_d)}{\sum_{i=1}^d \lambda_i(\boldsymbol{S}_d)}$. Combining inequalities (21), (22), (23), we obtain

$$1 \geq \frac{\mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right)}{\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right)} \geq 1 - \min\left\{\frac{d\cdot \mathrm{c}_1}{k}, \frac{d\cdot \mathrm{c}_2}{m}\right\}. \tag{24}$$

Since $\boldsymbol{S}_d$ is positive semi-definite, we have

$$\begin{aligned}
\mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d \boldsymbol{W}_m\right) &= \mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) + \mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^c \boldsymbol{W}_m\right) \\
&\geq \mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right).
\end{aligned} \tag{25}$$

Combining inequalities (24) and (25), we obtain $\varepsilon \leq \min\{(d\cdot \mathrm{c}_1/k), (d\cdot \mathrm{c}_2/m)\}$. Moreover, according to Theorem 4.3.53 [2], we have

$$\mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d \boldsymbol{W}_m\right) \geq \sum_{i=d-m+1}^{d} \lambda_i(\boldsymbol{S}_d) \geq m\cdot \lambda_d(\boldsymbol{S}_d). \tag{26}$$

Let $\kappa = \lambda_1(\boldsymbol{S}_d)/\lambda_d(\boldsymbol{S}_d)$, then by Ky Fan's Theorem [6], we obtain

$$\begin{aligned}\operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) &\leq \max_{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m}} \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d \boldsymbol{W}\right) = \sum_{i=1}^{m} \lambda_i(\boldsymbol{S}_d) \\ &\leq m \cdot \lambda_1(\boldsymbol{S}_d) = m \cdot \kappa \cdot \lambda_d(\boldsymbol{S}_d). \end{aligned} \tag{27}$$

Combing inequalities (26) and (27), we obtain $\varepsilon \leq 1 - \kappa^{-1}$. Furthermore, we have

$$\begin{aligned}\operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d \boldsymbol{W}_m\right) &\geq \operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) \\ &\geq \frac{k}{d} \operatorname{Tr}\left(\boldsymbol{S}_d^m\right) \geq \frac{k}{d} \sum_{i=1}^{m} \lambda_i(\boldsymbol{S}_d). \end{aligned} \tag{28}$$

According to inequality (27), we have

$$\operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) \leq \sum_{i=1}^{m} \lambda_i(\boldsymbol{S}_d). \tag{29}$$

Combing inequalities (28) and (29), we obtain $\varepsilon \leq 1 - k/d$. In summary, we conclude that $\varepsilon \leq \min\{(d \cdot \mathrm{c}_1/k), (d \cdot \mathrm{c}_2/m), 1 - \kappa^{-1}, 1 - k/d\}$. This completes the proof. □

References

[1] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Traces and Emergence of Nonlinear Programming*. Springer, 2014, pp. 247–258.

[2] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd ed. Cambridge; New York: Cambridge University Press, 2013.

[3] L. Tian, F. Nie, and X. Li, "Learning feature sparse principal components," *arXiv preprint arXiv:1904.10155*, 2019.

[4] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[5] F. Nie, D. Wu, R. Wang, and X. Li, "Truncated robust principle component analysis with a general optimization framework," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1081–1097, 2022.

[6] K. Fan, "On a theorem of weyl concerning eigenvalues of linear transformations: II," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 1, p. 31, 1950.