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
$\text{Tr}(\boldsymbol{X})$	Trace of square matrix $\mathbf{X}\in\mathbb{R}^{n\times n}$
$\text{rank}(\boldsymbol{X})$	Rank of matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$
$\boldsymbol{x}_i$	The i-th column of matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$
$\boldsymbol{x}^i$	The i-th row 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
$\|\boldsymbol{X}\|_F=\sqrt{\text{Tr}(\boldsymbol{X}^\top\boldsymbol{X})}$	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. The Workflow of SPDFS

The workflow of SPDFS is illustrated in Fig. 1.

III. Clarified Expression of Eq. (14)

For clarity, we present a more detailed and explicit derivation of Eq. (14) below.

$$
\begin{aligned}
&\sum_{i=1}^n\sum_{j=1}^c y_{ij}^r\|\boldsymbol{W}^\top\boldsymbol{x}_i-\boldsymbol{W}^\top\boldsymbol{u}_j\|_2^2\\
=&\sum_{i=1}^n\sum_{j=1}^c f_{ij}\left(\boldsymbol{x}_i^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{x}_i-2\boldsymbol{x}_i^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{u}_j+\boldsymbol{u}_j^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{u}_j\right)\\
=&\text{Tr}\left(\boldsymbol{W}^\top\boldsymbol{X}\text{diag}\left(\boldsymbol{F}\mathbf{1}\right)\boldsymbol{X}^\top\boldsymbol{W}\right)-\sum_{j=1}^c\frac{\boldsymbol{f}_j^\top\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{X}\boldsymbol{f}_j}{\boldsymbol{f}_j^\top\mathbf{1}}\\
=&\text{Tr}\left(\boldsymbol{W}^\top\boldsymbol{X}\left(\text{diag}\left(\boldsymbol{F}\mathbf{1}\right)-\boldsymbol{F}\text{diag}\left(\boldsymbol{F}^\top\mathbf{1}\right)^{-1}\boldsymbol{F}^\top\right)\boldsymbol{X}^\top\boldsymbol{W}\right)\\
=&\text{Tr}\left(\boldsymbol{W}^\top\boldsymbol{X}\left(\boldsymbol{D}-\boldsymbol{G}\right)\boldsymbol{X}^\top\boldsymbol{W}\right)\\
=&\text{Tr}\left(\boldsymbol{W}^\top\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top\boldsymbol{W}\right)\\
=&\text{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right),
\end{aligned}\tag{1}
$$

where $\boldsymbol{S}_m=\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top$, and $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{G}=\text{diag}\left(\boldsymbol{F}\mathbf{1}\right)-\boldsymbol{F}\text{diag}\left(\boldsymbol{F}^\top\mathbf{1}\right)^{-1}\boldsymbol{F}^\top$ is indeed the Laplacian matrix in graph theory. To see why, we analyse its two components, $\boldsymbol{D}$ and $\boldsymbol{G}$, separately. First, $\boldsymbol{G}=\boldsymbol{F}\text{diag}\left(\boldsymbol{F}^\top\mathbf{1}\right)^{-1}\boldsymbol{F}^\top$ serves as a normalized similarity matrix, capturing the pairwise similarity among samples while incorporating class importance normalization. Second, given the definition of the degree matrix, $\boldsymbol{D}=\text{diag}\left(\boldsymbol{G}\mathbf{1}\right)$. By direct derivation, we have $\boldsymbol{D}=\text{diag}\left(\boldsymbol{F}\mathbf{1}\right)$.

IV. Practical and Efficient Choice of γ

In problem (16), $\boldsymbol{S}_d=\gamma\boldsymbol{I}-\boldsymbol{S}_o$, where $\boldsymbol{S}_o=\boldsymbol{S}_m-\alpha\boldsymbol{S}_t$. γ is large enough to ensure $\boldsymbol{S}_d$ is positive semi-definite. Theoretically, γ can be set to the largest eigenvalue of $\boldsymbol{S}_o$, i.e., $\lambda_{\max}(\boldsymbol{S}_o)$. However, computing $\lambda_{\max}(\boldsymbol{S}_o)$ via eigenvalue decomposition is computationally expensive. Instead, for the square matrix $\boldsymbol{S}_o$, the 1-norm $\|\boldsymbol{S}_o\|_1$ and infinity norm $\|\boldsymbol{S}_o\|_\infty$ provide efficient upper bounds on $\lambda_{\max}(\boldsymbol{S}_o)$ without requiring explicit eigenvalue computation [2]. Since $\boldsymbol{S}_o$ is symmetric, $\|\boldsymbol{S}_o\|_1=\|\boldsymbol{S}_o\|_\infty$, making them equivalent choices for γ and ensuring computational efficiency.

V. An Example of Matrix A

To clarify the description of matrix $\boldsymbol{A}\in\{0,1\}^{d\times k}$ in Section IV-A1, 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 $\boldsymbol{A}^\top\mathbf{1}_d=\mathbf{1}_k$.

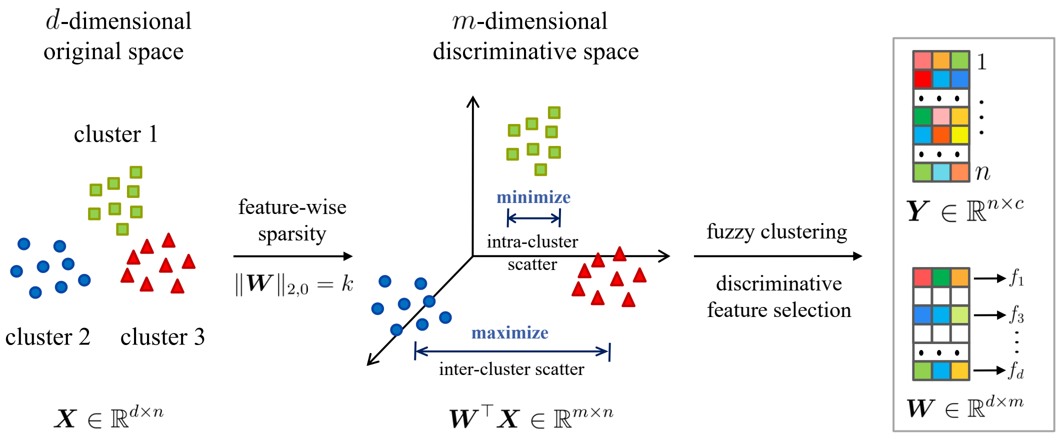

Fig. 1: Illustration of the SPDFS workflow. Guided by the principle of supervised LDA, SPDFS jointly performs fuzzy c-means membership learning $\boldsymbol{Y} \in \mathbb{R}^{n\times c}$ and PCA projection learning $\boldsymbol{W} \in \mathbb{R}^{d\times m}$ under an $\ell_{2,0}$-norm constraint $\|\boldsymbol{W}\|_{2,0} = k$ for feature-wise sparsity, enabling discriminative feature selection in an unsupervised manner.

Algorithm 3: Solve Problem (6).

Input: $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, $\boldsymbol{S}_d \in \mathbb{R}^{d\times d}$, d, k, m, r.
Initialization: Initialize $\boldsymbol{Y}_0$ and $\boldsymbol{M}_0$ by Eq. (5), and initialize $\boldsymbol{W}_0$ randomly.
while *not converge* **do**
  Update α by Eq. (9).
  while *not converge* **do**
    Update $\boldsymbol{M}$ by Eq. (11).
    Update $\boldsymbol{W}$ by Algorithm 2.
    Update $\boldsymbol{Y}$ by Eq. (26).
Output: $\boldsymbol{W} \in \mathbb{R}^{d\times m}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times c}$, $\boldsymbol{M} \in \mathbb{R}^{m\times c}$.

VI. Relationship Between $\boldsymbol{S}_m$ and $\boldsymbol{S}_w$

In problem (36), $\boldsymbol{S}_w$ is the intra-cluster scatter matrix in LDA, given by $\boldsymbol{S}_w = \sum_{j=1}^{c}\sum_{y_{ij}=1}\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j\|_2^2 = \left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{Y}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^\top\right)\left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{Y}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^\top\right)^\top = \boldsymbol{X}\left(\boldsymbol{I} - \boldsymbol{Y}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^\top\right)\boldsymbol{X}^\top$. In fact, this structure can be directly observed from Eq. (1). Specifically, when $r = 1$, we have $f_{ij} = y_{ij}^r = y_{ij}$, leading to $\boldsymbol{S}_m = \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^\top = \boldsymbol{X}\left(\mathrm{diag}\left(\boldsymbol{Y}\mathbf{1}\right) - \boldsymbol{Y}\mathrm{diag}\left(\boldsymbol{Y}^\top\mathbf{1}\right)^{-1}\boldsymbol{Y}^\top\right)\boldsymbol{X}^\top$. Since $\boldsymbol{Y}$ satisfies $\sum_{j=1}^{c} y_{ij} = 1$ and $y_{ij} \in \{0,1\}$, it follows that $\boldsymbol{S}_m = \boldsymbol{X}\left(\boldsymbol{I} - \boldsymbol{Y}\left(\boldsymbol{Y}^\top\boldsymbol{Y}\right)^{-1}\boldsymbol{Y}^\top\right)\boldsymbol{X}^\top = \boldsymbol{S}_w$. This reveals the relationship between $\boldsymbol{S}_m$ and $\boldsymbol{S}_w$. That is, when $r = 1$, we have $\boldsymbol{S}_m = \boldsymbol{S}_w$.

VII. Correction to Algorithm 3

The pseudocode for Algorithm 3 in the main text of the published article inadvertently missed a line, specifically the optimization step for variable $\boldsymbol{M}$. The complete Algorithm 3 is provided here as a supplement.

VIII. Correction to The Proof of Theorem 6

The proof of Theorem 6 in the main text has been revised for clarity and completeness. The updated version presented here provides a more accurate and complete presentation of the proof.

Proof. We begin with problem (31), which can be equivalently expressed as follows:

$$\min_{\mathrm{Tr}(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W})=\mathrm{C}} \frac{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right)}{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right)}. \tag{2}$$

Observe that this trace ratio formulation can be rewritten as: $\frac{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right)}{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right)} = \frac{\sum_{i=1}^{m}\boldsymbol{w}_i^\top\boldsymbol{S}_m\boldsymbol{w}_i}{\sum_{i=1}^{m}\boldsymbol{w}_i^\top\boldsymbol{S}_t\boldsymbol{w}_i}$. Suppose that $\frac{\boldsymbol{w}_1^\top\boldsymbol{S}_m\boldsymbol{w}_1}{\boldsymbol{w}_1^\top\boldsymbol{S}_t\boldsymbol{w}_1}$ is the minimum among the set $\left\{\frac{\boldsymbol{w}_i^\top\boldsymbol{S}_m\boldsymbol{w}_i}{\boldsymbol{w}_i^\top\boldsymbol{S}_t\boldsymbol{w}_i}\right\}_{i=1}^{m}$. By Lemma 2, we have $\frac{\boldsymbol{w}_1^\top\boldsymbol{S}_m\boldsymbol{w}_1}{\boldsymbol{w}_1^\top\boldsymbol{S}_t\boldsymbol{w}_1} \leq \frac{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right)}{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right)}$. Since $\boldsymbol{w}_*$ is defined as $\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top\boldsymbol{S}_m\boldsymbol{w}}{\boldsymbol{w}^\top\boldsymbol{S}_t\boldsymbol{w}}$, it follows that for any $\boldsymbol{W}$, $\frac{\boldsymbol{w}_*^\top\boldsymbol{S}_m\boldsymbol{w}_*}{\boldsymbol{w}_*^\top\boldsymbol{S}_t\boldsymbol{w}_*} \leq \frac{\boldsymbol{w}_1^\top\boldsymbol{S}_m\boldsymbol{w}_1}{\boldsymbol{w}_1^\top\boldsymbol{S}_t\boldsymbol{w}_1} \leq \frac{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right)}{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right)}$. When each column of $\boldsymbol{W}$ is equal to $\boldsymbol{w}_*$, i.e., $\boldsymbol{w}_i = \boldsymbol{w}_*$ for all $i \in [1, m]$, the equality in $\frac{\boldsymbol{w}_*^\top\boldsymbol{S}_m\boldsymbol{w}_*}{\boldsymbol{w}_*^\top\boldsymbol{S}_t\boldsymbol{w}_*} \leq \frac{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right)}{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right)}$ holds. Thus, the minimum value of $\frac{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_m\boldsymbol{W}\right)}{\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right)}$ is achieved at $\frac{\boldsymbol{w}_*^\top\boldsymbol{S}_m\boldsymbol{w}_*}{\boldsymbol{w}_*^\top\boldsymbol{S}_t\boldsymbol{w}_*}$. To satisfy the constraint $\mathrm{Tr}\left(\boldsymbol{W}^\top\boldsymbol{S}_t\boldsymbol{W}\right) = \mathrm{C}$, an optimal solution to problem (31) is $\boldsymbol{W}_* = [\mathrm{c}_1\boldsymbol{w}_*, \mathrm{c}_2\boldsymbol{w}_*, \ldots, \mathrm{c}_m\boldsymbol{w}_*]$, which is a trivial solution since all columns are multiples of the same vector $\boldsymbol{w}_*$, making the rank of $\boldsymbol{W}_*$ at most 1, under the assumption that $\boldsymbol{w}_*$ is the unique solution. If $\boldsymbol{w}_*$ is not unique, then each column of the optimal $\boldsymbol{W}_*$ lies within the subspace spanned by the solutions of $\boldsymbol{w}_*$. Here, $\{\mathrm{c}_i\}_{i=1}^{m}$ are arbitrary constants chosen such that $\mathrm{Tr}\left(\boldsymbol{W}_*^\top\boldsymbol{S}_t\boldsymbol{W}_*\right) = \mathrm{C}$. This completes the proof. $\square$

IX. 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, $\forall\ \boldsymbol{x} \in \mathcal{S}$, 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 $\min_{\boldsymbol{x}\in\mathcal{S}}\left(h(\boldsymbol{x}) - \alpha_* p(\boldsymbol{x})\right) = 0$. Now, define the function $f(\alpha) = \min_{\boldsymbol{x}\in\mathcal{S}}\left(h(\boldsymbol{x}) - \alpha p(\boldsymbol{x})\right)$. Then, we have $f(\alpha_*) = 0$. This completes the proof. $\square$

X. 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. □

XI. The Proof of Remark 1

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

$$f(\boldsymbol{W}_t) = \mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right), \tag{3}$$

$$g\left(\boldsymbol{W}_t|\boldsymbol{W}_t\right) = \mathrm{Tr}\left(\boldsymbol{W}_t^\top \left(\boldsymbol{S}_d \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{S}_d\right) \boldsymbol{W}_t\right). \tag{4}$$

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:

$$\boldsymbol{\Upsilon} = \boldsymbol{Q}^\top \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{Q}, \tag{5}$$

$$\boldsymbol{\Psi} = \boldsymbol{Q}^\top \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{Q}. \tag{6}$$

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

$$\begin{aligned} g\left(\boldsymbol{W}|\boldsymbol{W}_t\right) &= \mathrm{Tr}\left(\boldsymbol{W}^\top \left(\boldsymbol{S}_d \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{S}_d\right) \boldsymbol{W}\right) \\ &= \mathrm{Tr}\left(\boldsymbol{\Upsilon}\boldsymbol{\Psi}\right). \end{aligned} \tag{7}$$

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

$$\mathrm{Tr}\left(\boldsymbol{\Upsilon}\boldsymbol{\Psi}\right) \leq \sum_{i=1}^{d} \lambda_i\left(\boldsymbol{\Upsilon}\right) \lambda_i\left(\boldsymbol{\Psi}\right) \leq \sum_{i=1}^{m} \lambda_i\left(\boldsymbol{\Psi}\right). \tag{8}$$

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

XII. 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:

$$f(\widetilde{\boldsymbol{W}}_{t+1}) \geq g(\widetilde{\boldsymbol{W}}_{t+1}|\boldsymbol{W}_t) \geq g(\boldsymbol{W}_t|\boldsymbol{W}_t) = f(\boldsymbol{W}_t). \tag{9}$$

According to Eq. (3) and inequality (9), we have

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

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

$$\begin{aligned} \mathrm{Tr}\left(\widetilde{\boldsymbol{W}}_{t+1}^\top \boldsymbol{S}_d \widetilde{\boldsymbol{W}}_{t+1}\right) &= \mathrm{Tr}\left(\widetilde{\boldsymbol{B}}_{t+1}^\top \boldsymbol{A}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{A}_{t+1} \widetilde{\boldsymbol{B}}_{t+1}\right) \\ &\leq \mathrm{Tr}\left(\boldsymbol{B}_{t+1}^\top \boldsymbol{A}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{A}_{t+1} \boldsymbol{B}_{t+1}\right) \\ &= \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right). \end{aligned} \tag{11}$$

According to inequalities (10) and (11), we have

$$\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \leq \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right). \tag{12}$$

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 $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \neq \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$. This result demonstrates the ascent property of Algorithm 2, namely, $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) < \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$.

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 $\mathrm{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \mathrm{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$. Then the equality in inequality (8) 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

$$\boldsymbol{Q}^\top \boldsymbol{W}_t \left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)^\dagger \boldsymbol{W}_t^\top \boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1} \boldsymbol{W}_t^\top \boldsymbol{Q} = \boldsymbol{\Phi}_t \boldsymbol{\Phi}_t^\top, \tag{13}$$

$$\boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t = \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\Omega}_t \boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{I}_{m \times m}. \tag{14}$$

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

$$\begin{aligned} \boldsymbol{Q}^\top \boldsymbol{W}_{t+1} \boldsymbol{W}_{t+1}^\top \boldsymbol{Q} &= \boldsymbol{\Phi}_t \boldsymbol{\Omega}_{t+1} \boldsymbol{\Phi}_t^\top \\ &= \boldsymbol{Q}^\top \boldsymbol{W}_t \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{\Omega}_{t+1} \boldsymbol{\Omega}_t^{-1/2} \boldsymbol{W}_t^\top \boldsymbol{Q}. \end{aligned} \tag{15}$$

Based on Eq. (15), we have

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

$$= \mathbf{\Omega}_t\mathbf{\Omega}_{t+1}. \tag{16}$$

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

$$\begin{aligned} &\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)^\dagger \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ =\ &\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1}\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ =\ &\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1}\boldsymbol{W}_{t+1}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_t\right) \\ =\ &\operatorname{Tr}\left(\boldsymbol{Q}^\top \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1}\boldsymbol{W}_{t+1}^\top \boldsymbol{Q}\boldsymbol{Q}^\top \boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{Q}\right) \\ =\ &\operatorname{Tr}\left(\mathbf{\Upsilon}_{t+1}\mathbf{\Psi}_t\right). \end{aligned} \tag{17}$$

From Eq. (17), inequality (8), and Theorem 4.3.53 [2], we obtain

$$\begin{aligned} &\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1}\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ \leq\ &\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \operatorname{Tr}\left(\mathbf{\Omega}_t\right). \end{aligned} \tag{18}$$

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

$$\begin{aligned} &\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1}\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \\ =\ &\operatorname{Tr}\left(\mathbf{\Gamma}\mathbf{\mho}\right) \geq \sum_{i=1}^{m} \lambda_i(\mathbf{\Gamma})\lambda_{m-i+1}(\mathbf{\mho}). \end{aligned} \tag{19}$$

Note that $\lambda_{m-i+1}(\mathbf{\mho}) = \lambda_i(\mathbf{\Omega}_{t+1})^{-1}$, and by Eq. (16) and Theorem 1.3.22 [2], we get

$$\sum_{i=1}^{m} \lambda_i(\mathbf{\Gamma})\lambda_{m-i+1}(\mathbf{\mho}) = \sum_{i=1}^{m} \frac{\lambda_i(\mathbf{\Omega}_t\mathbf{\Omega}_{t+1})}{\lambda_i(\mathbf{\Omega}_{t+1})} = \operatorname{Tr}\left(\mathbf{\Omega}_t\right). \tag{20}$$

Combing inequalities (18), (19) and Eq. (20), we conclude that $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1}\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \operatorname{Tr}\left(\mathbf{\Omega}_t\right)$. According to Theorem 4.3.53 [2], the equality in inequality (19) implies that $\mathbf{\Gamma}$ and $\mathbf{\mho}$ are simultaneously diagonalizable. From Eq. (16), we know that $\mathbf{\Gamma} = \boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} = \mathbf{\Omega}_t\mathbf{\Omega}_{t+1}$, which implies that $\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t = \mathbf{\Omega}_t^{1/2}\mathbf{\Omega}_{t+1}^{1/2}$ is diagonal. Recall that $\mathbf{\Phi}_t = \boldsymbol{Q}^\top \boldsymbol{W}_t \mathbf{\Omega}_t^{-1/2}$ and $\mathbf{\Phi}_{t+1} = \boldsymbol{Q}^\top \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1/2}$, then we have

$$\begin{aligned} &\mathbf{\Phi}_t^\top \mathbf{\Phi}_t = \mathbf{\Omega}_t^{-1/2}\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t \mathbf{\Omega}_t^{-1/2} = \boldsymbol{I}_{m\times m}, \\ &\mathbf{\Phi}_{t+1}^\top \mathbf{\Phi}_{t+1} = \mathbf{\Omega}_{t+1}^{-1/2}\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1} \mathbf{\Omega}_{t+1}^{-1/2} = \boldsymbol{I}_{m\times m}, \\ &\mathbf{\Phi}_{t+1}^\top \mathbf{\Phi}_t = \mathbf{\Omega}_{t+1}^{-1/2}\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_t \mathbf{\Omega}_t^{-1/2} = \boldsymbol{I}_{m\times m}. \end{aligned} \tag{21}$$

Thus, we conclude that $\mathbf{\Phi}_t = \mathbf{\Phi}_{t+1}$, which implies that $\boldsymbol{S}_d \boldsymbol{W}_t \mathbf{\Omega}_t^{-1/2} = \boldsymbol{S}_d \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1/2}$. Since $\boldsymbol{S}_d$ is full rank, it follows that $\boldsymbol{W}_t \mathbf{\Omega}_t^{-1/2} = \boldsymbol{W}_{t+1}\mathbf{\Omega}_{t+1}^{-1/2}$. Therefore, we conclude that $\boldsymbol{A}_t = \boldsymbol{A}_{t+1}$, since the operation $\mathbf{\Omega}^{-1/2}$ 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 $\{\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)\}_{t\in\mathcal{Z}^+}$. Note that $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right)$ is upper bounded by $\operatorname{Tr}\left(\boldsymbol{S}_d\right)$. 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 $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) < \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$. Therefore, by the contrapositive, if the objective value converges, i.e., $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) = \operatorname{Tr}\left(\boldsymbol{W}_{t+1}^\top \boldsymbol{S}_d \boldsymbol{W}_{t+1}\right)$, then 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 $\operatorname{Tr}\left(\boldsymbol{W}_t^\top \boldsymbol{S}_d \boldsymbol{W}_t\right) \to \operatorname{Tr}\left(\widehat{\boldsymbol{W}}^\top \boldsymbol{S}_d \widehat{\boldsymbol{W}}\right)$. This completes the proof. □

XIII. 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:

$$\begin{aligned} &\operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) \\ =&\max_{\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{m\times m},\ \|\boldsymbol{W}\|_{2,0}=k} \operatorname{Tr}\left(\boldsymbol{W}^\top\left(\boldsymbol{S}_d^m + \boldsymbol{S}_d^c\right)\boldsymbol{W}\right) \\ \leq&\max_{\substack{\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{m\times m},\\ \|\boldsymbol{W}\|_{2,0}=k}} \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^m \boldsymbol{W}\right) + \max_{\substack{\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{m\times m},\\ \|\boldsymbol{W}\|_{2,0}=k}} \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^c \boldsymbol{W}\right) \\ \leq&\operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) + \max_{\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{m\times m}} \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^c \boldsymbol{W}\right) \\ \leq&\operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) + \sum_{i=m+1}^{2m}\lambda_i(\boldsymbol{S}_d), \end{aligned} \tag{22}$$

from which it follows that

$$\frac{\operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right)}{\operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right)} \geq 1 - \frac{\sum_{i=m+1}^{2m}\lambda_i(\boldsymbol{S}_d)}{\operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right)}. \tag{23}$$

Furthermore, we have the following two observations:

$$\begin{aligned} \operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) &\overset{(a)}{\geq} \max_{\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{m\times m},\ \|\boldsymbol{W}\|_{2,0}=m} \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d \boldsymbol{W}\right) \\ &\overset{(b)}{\geq} \frac{m}{d}\operatorname{Tr}\left(\boldsymbol{S}_d\right) = \frac{m}{d}\sum_{i=1}^{d}\lambda_i(\boldsymbol{S}_d), \end{aligned} \tag{24}$$

$$\begin{aligned} \operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) &\overset{(c)}{\geq} \max_{\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{m\times m},\ \|\boldsymbol{W}\|_{2,0}=k} \operatorname{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d^m \boldsymbol{W}\right) \\ &\overset{(d)}{\geq} \frac{k}{d}\sum_{i=1}^{d}\lambda_i(\boldsymbol{S}_d^m) = \frac{k}{d}\sum_{i=1}^{m}\lambda_i(\boldsymbol{S}_d), \end{aligned} \tag{25}$$

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 $\operatorname{rank}(\boldsymbol{S}_d^m) = m$, and thus the problem on the left side of (d) also achieves its global optimum via Algorithm 2.

Let $\mathrm{z} = \min\{\operatorname{rank}\left(\boldsymbol{S}_d\right), 2m\}$, $\mathrm{c}_1 = \frac{\sum_{i=m+1}^{z}\lambda_i(\boldsymbol{S}_d)}{\sum_{i=1}^{m}\lambda_i(\boldsymbol{S}_d)}$, and $\mathrm{c}_2 = \frac{\sum_{i=m+1}^{z}\lambda_i(\boldsymbol{S}_d)}{\sum_{i=1}^{d}\lambda_i(\boldsymbol{S}_d)}$. Combining inequalities (23), (24), (25), we obtain

$$1 \geq \frac{\operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right)}{\operatorname{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right)} \geq 1 - \min\left\{\frac{d\cdot \mathrm{c}_1}{k}, \frac{d\cdot \mathrm{c}_2}{m}\right\}. \tag{26}$$

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

$$\begin{aligned} \operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d \boldsymbol{W}_m\right) &= \operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) + \operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^c \boldsymbol{W}_m\right) \\ &\geq \operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right). \end{aligned} \tag{27}$$

Combining inequalities (26) and (27), 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

$$\operatorname{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d \boldsymbol{W}_m\right) \geq \sum_{i=d-m+1}^{d}\lambda_i(\boldsymbol{S}_d) \geq m\cdot\lambda_d(\boldsymbol{S}_d). \tag{28}$$

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

$$\begin{aligned}\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) \le \max_{\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{m\times m}} \mathrm{Tr}\left(\boldsymbol{W}^\top \boldsymbol{S}_d \boldsymbol{W}\right) = \sum_{i=1}^{m} \lambda_i(\boldsymbol{S}_d) \\ \le m \cdot \lambda_1(\boldsymbol{S}_d) = m \cdot \kappa \cdot \lambda_d(\boldsymbol{S}_d). \end{aligned} \tag{29}$$

Combing inequalities (28) and (29), we obtain $\varepsilon \le 1 - \kappa^{-1}$. Furthermore, we have

$$\begin{aligned}\mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d \boldsymbol{W}_m\right) &\ge \mathrm{Tr}\left(\boldsymbol{W}_m^\top \boldsymbol{S}_d^m \boldsymbol{W}_m\right) \\ &\ge \frac{k}{d}\mathrm{Tr}\left(\boldsymbol{S}_d^m\right) \ge \frac{k}{d}\sum_{i=1}^{m} \lambda_i(\boldsymbol{S}_d). \end{aligned} \tag{30}$$

According to inequality (29), we have

$$\mathrm{Tr}\left(\boldsymbol{W}_*^\top \boldsymbol{S}_d \boldsymbol{W}_*\right) \le \sum_{i=1}^{m} \lambda_i(\boldsymbol{S}_d). \tag{31}$$

Combing inequalities (30) and (31), we obtain $\varepsilon \le 1 - k/d$. In summary, we conclude that $\varepsilon \le \min\{(d \cdot \mathrm{c}_1/k), (d \cdot \mathrm{c}_2/m), 1 - \kappa^{-1}, 1 - k/d\}$. 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.