Unsupervised Simultaneous Orthogonal Basis Clustering Feature Selection - Supplementary Material

Dongyoon Han and Junmo Kim
School of Electrical Engineering, KAIST, South Korea
{dyhan, junmo.kim}@kaist.ac.kr

Algorithm 1: $\mathbf{E}$, $\mathbf{F}$ updates algorithm

Input: $\mathbf{F}_t$, $\mathbf{W}_t$ and $\mathbf{B}_t$; Parameter: γ
Initialization: $s = 0$ and $\mathbf{F}'_s = \mathbf{F}_t$
repeat
  Update $\mathbf{E}'_{s+1} = \mathbf{V_E}\mathbf{I}_{n,c}\mathbf{U_E}^T$ by (4) where $\mathbf{B}^T\mathbf{W}_t^T\mathbf{X}_t + \gamma\mathbf{F}'^T_s = \mathbf{U_E}\mathbf{\Sigma_E}\mathbf{V_E}^T$;
  Update $\mathbf{F}'_{s+1} = \frac{\mathbf{E}'_{s+1} + |\mathbf{E}'_{s+1}|}{2}$ by (5);
  $s = s + 1$;
until $\|\triangle J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_s)\| \le \epsilon \;\; or \;\; s \le S$;
Output: $\mathbf{E}_{t+1} = \mathbf{E}'_s$, $\mathbf{F}_{t+1} = \mathbf{F}'_s$

Algorithm 2: SOCFS

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$; Parameters: λ, γ
Initialization: $t = 0$, $\mathbf{D}_t = \mathbf{I}$ and $\mathbf{B}_t, \mathbf{E}_t$
repeat
  Update $\mathbf{E}_{t+1}$ and $\mathbf{F}_{t+1}$ by Algorithm 1;
  Update $\mathbf{W}_{t+1} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{D}_t)^{-1}\mathbf{X}\mathbf{E}_{t+1}\mathbf{B}_t^T$ by (2);
  Update $\mathbf{B}_{t+1} = \mathbf{V_B}\mathbf{I}_{m,c}\mathbf{U_B}^T$ by (3) where $\mathbf{E}_{t+1}^T\mathbf{X}^T\mathbf{W}_{t+1} = \mathbf{U_B}\mathbf{\Sigma_B}\mathbf{V_B}^T$;
  Update the i-th diagonal elements of the diagonal matrix $\mathbf{D}_{t+1}$ with $\frac{1}{2\|\mathbf{w}_{t+1}^i\|_2}$;
  $t = t + 1$;
until $\|\triangle J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_t, \mathbf{F}_t)\| \le \epsilon \;\; or \;\; t \le T$;
Output: Features are selected corresponding to the largest values of $\|\mathbf{w_t^i}\|$, $i = 1 \ldots d$, which are sorted by descending order.

1. Preliminaries

1.1. The Reformulated Objective Function

$$\min_{\mathbf{W},\mathbf{B},\mathbf{E},\mathbf{F}} \; \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} + \gamma\|\mathbf{F} - \mathbf{E}\|_F^2$$
$$s.t. \;\; \mathbf{B}^T\mathbf{B} = \mathbf{I}, \; \mathbf{E}^T\mathbf{E} = \mathbf{I}, \; \mathbf{F} \ge \mathbf{0}. \quad (1)$$

1.2. Update Rules

W update:

$$\mathbf{W} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{D})^{-1}\mathbf{X}\mathbf{E}\mathbf{B}^T. \quad (2)$$

B update:

$$\mathbf{B} = \mathbf{V_B}\mathbf{I}_{m,c}\mathbf{U_B}^T, \quad (3)$$

where $\mathbf{U_B}$ and $\mathbf{V_B}$ are the left and right eigenvectors of $\mathbf{E}^T\mathbf{X}^T\mathbf{W}$ computed by SVD, respectively.

$\mathbf{E}$, $\mathbf{F}$ update:

$$\mathbf{E} = \mathbf{V_E}\mathbf{I}_{n,c}\mathbf{U_E}^T, \quad (4)$$

where $\mathbf{U_E}$ and $\mathbf{V_E}$ are the left and right eigenvectors of $\mathbf{B}^T\mathbf{W}^T\mathbf{X} + \gamma\mathbf{F}^T$ computed by SVD, respectively.

$$\mathbf{F} = \frac{1}{2}(\mathbf{E} + |\mathbf{E}|). \quad (5)$$

1.3. Algorithms

The optimization algorithm containing the $\mathbf{E}$ and $\mathbf{F}$ update rules is summarized in Algorithm 1. The overall proposed optimization algorithm of SOCFS is also presented in Algorithm 2.

2. Convergence Analysis

We prove the convergence of the proposed optimization algorithm with monotonic decrease at every iteration. We denote the objective function in problem (1) as $J(\mathbf{W}, \mathbf{B}, \mathbf{E}, \mathbf{F})$ for convenience.

Theorem 1. *$J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_s) \triangleq J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}'_s, \mathbf{F}'_s)$ monotonically decreases due to E, F updates in Algorithm 1.*

Proof. For the $\mathbf{F}'$ update from by (5), we have

$$\mathbf{F}'_{s+1} = \underset{\mathbf{F}':\mathbf{F}'\succeq\mathbf{0}}{\arg\min} \; \|\mathbf{F}' - \mathbf{E}'_s\|_F^2 = \underset{\mathbf{F}':\mathbf{F}'\succeq\mathbf{0}}{\arg\min} \; J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}')$$
$$\Longrightarrow J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_{s+1}) \le J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_s). \quad (6)$$

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

Similarly, for the $\mathbf{E}'$ update by (4), we have

$$\begin{aligned} \mathbf{E}'_{s+1} &= \underset{\mathbf{E}':\mathbf{E}'^T\mathbf{E}'=\mathbf{I}}{\arg\min} \ \|\mathbf{B}_t\mathbf{E}'^T - \mathbf{W}_t^T\mathbf{X}\|_F^2 + \gamma\|\mathbf{E}' - \mathbf{F}'_{s+1}\|_F^2 \\ &= \underset{\mathbf{E}':\mathbf{E}'^T\mathbf{E}'=\mathbf{I}}{\arg\min} \ J_{EF}^{(t)}(\mathbf{E}', \mathbf{F}'_{s+1}) \\ &\Longrightarrow J_{EF}^{(t)}(\mathbf{E}'_{s+1}, \mathbf{F}'_{s+1}) \leq J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_{s+1}). \end{aligned} \tag{7}$$

By combining (6) and (7), we finally obtain

$$J_{EF}^{(t)}(\mathbf{E}'_{s+1}, \mathbf{F}'_{s+1}) \leq J_{EF}^{(t)}(\mathbf{E}'_{s+1}, \mathbf{F}'_s) \leq J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_s).$$

Thus $J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_s)$ monotonically decreases by the update rules (4) and (5) in Algorithm 1. We also notice that, since $J_{EF}^{(t)}(\mathbf{E}'_s, \mathbf{F}'_s)$ is convex in each variable, the algorithm must converge. □

Theorem 2. *$J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_t, \mathbf{F}_t)$ monotonically decreases due to the update rules in Algorithm 2.*

Proof. For the $\mathbf{E}$ and $\mathbf{F}$ updates, $\mathbf{E}_{t+1}$ and $\mathbf{F}_{t+1}$ are updated at the same time by Algorithm 1, so that we have

$$J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_{t+1}, \mathbf{F}_{t+1}) \leq J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_t, \mathbf{F}_t). \tag{8}$$

For the $\mathbf{W}$ update by (2), which follows the theorem in [1] closely, $\mathbf{W}_{t+1}$ is also the solution of the following problem with fixed $\mathbf{D}_t$ as

$$\mathbf{W}_{t+1} = \underset{\mathbf{W}}{\arg\min} \ \|\mathbf{W}_t^T\mathbf{X} - \mathbf{B}_t\mathbf{E}_t^T\|_F^2 + \lambda \operatorname{tr}(\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t).$$

This implies that

$$\begin{aligned} &\|\mathbf{W}_{t+1}^T\mathbf{X} - \mathbf{B}_t\mathbf{E}_t^T\|_F^2 + \lambda \operatorname{tr}(\mathbf{W}_{t+1}^T\mathbf{D}_t\mathbf{W}_{t+1}) \\ &\quad \leq \|\mathbf{W}_t^T\mathbf{X} - \mathbf{B}_t\mathbf{E}_t^T\|_F^2 + \lambda \operatorname{tr}(\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t). \end{aligned} \tag{9}$$

And then according to the lemma in [1] with $\mathbf{u} = \mathbf{w}_{t+1}^i$, $\mathbf{u}_t = \mathbf{w}_t^i$ and summation over all rows, we have

$$\sum_{i=1}^d \left(\|\mathbf{w}_{t+1}^i\|_2 - \frac{\|\mathbf{w}_{t+1}^i\|_2^2}{2\|\mathbf{w}_t^i\|_2} \right) \leq \sum_{i=1}^d \left(\|\mathbf{w}_t^i\|_2 - \frac{\|\mathbf{w}_t^i\|_2^2}{2\|\mathbf{w}_t^i\|_2} \right).$$

We rewrite the inequality as

$$\begin{aligned} &\|\mathbf{W}_{t+1}\|_{2,1} - \operatorname{tr}(\mathbf{W}_{t+1}^T\mathbf{D}_t\mathbf{W}_{t+1}) \\ &\qquad \leq \|\mathbf{W}_t\|_{2,1} - \operatorname{tr}(\mathbf{W}_t^T\mathbf{D}_t\mathbf{W}_t). \end{aligned} \tag{10}$$

By combining (9) and (10), we finally obtain

$$J(\mathbf{W}_{t+1}, \mathbf{B}_t, \mathbf{E}_{t+1}, \mathbf{F}_{t+1}) \leq J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_{t+1}, \mathbf{F}_{t+1}). \tag{11}$$

For the $\mathbf{B}$ update by (3), we have

$$\begin{aligned} \mathbf{B}_{t+1} &= \underset{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}}{\arg\min} \ \|\mathbf{E}_{t+1}\mathbf{B}^T - \mathbf{X}^T\mathbf{W}_{t+1}\|_F^2 \\ &= \underset{\mathbf{B}:\mathbf{B}^T\mathbf{B}=\mathbf{I}}{\arg\min} \ J_B(\mathbf{W}_{t+1}, \mathbf{B}, \mathbf{E}_{t+1}, \mathbf{F}_{t+1}). \end{aligned} \tag{12}$$

This implies that

$$J(\mathbf{W}_{t+1}, \mathbf{B}_{t+1}, \mathbf{E}_{t+1}, \mathbf{F}_{t+1}) \leq J(\mathbf{W}_{t+1}, \mathbf{B}_t, \mathbf{E}_{t+1}, \mathbf{F}_{t+1}). \tag{13}$$

From (8), (11), and (13), each update rule monotonically decreases the objective function at every iteration. We also notice that, since $J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_t, \mathbf{F}_t)$ is convex in each variable, the algorithm with the update rules must converge. □

References

[1] F. Nie, H. Huang, X. Cai, and C. H. Ding. Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010. 2