Unsupervised Simultaneous Orthogonal Basis Clustering Feature Selection

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

Abstract

In this paper, we propose a novel unsupervised feature selection method: Simultaneous Orthogonal basis Clustering Feature Selection (SOCFS). To perform feature selection on unlabeled data effectively, a regularized regression-based formulation with a new type of target matrix is designed. The target matrix captures latent cluster centers of the projected data points by performing orthogonal basis clustering, and then guides the projection matrix to select discriminative features. Unlike the recent unsupervised feature selection methods, SOCFS does not explicitly use the pre-computed local structure information for data points represented as additional terms of their objective functions, but directly computes latent cluster information by the target matrix conducting orthogonal basis clustering in a single unified term of the proposed objective function. It turns out that the proposed objective function can be minimized by a simple optimization algorithm. Experimental results demonstrate the effectiveness of SOCFS achieving the state-of-the-art results with diverse real world datasets.

1. Introduction

In recent years, high-dimensional features have been widely used as inputs of several learning tasks. However, if one uses these high-dimensional features directly, unimportant features to learning tasks lead to heavy computational complexity and critical degradation of the performance of the learning tasks. Most high-dimensional features contain unimportant features that are largely categorized as 1) redundant features correlated to others or 2) noisy features.

To deal with this problem, it is natural to filter out those unimportant features from the original features. Typically two kind of methods (feature extraction and feature selection) have been developed. Unlike feature selection methods, feature extraction methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) change the feature space to construct renewed features, so the values of each feature change. Therefore, feature selection has an advantage to select discriminative features while preserving their values.

A number of feature selection methods have been proposed and classified as supervised and unsupervised feature selection methods. Supervised feature selection methods such as [6, 15, 24, 32, 21, 20] improve results of the learning tasks by exploiting label information. However, label information is usually expensive in practice, so unsupervised feature selection is of more practical importance. Many unsupervised feature selection methods such as [13, 32, 21, 3, 30, 18, 23, 28] have been proposed to deal with the unlabeled data. One type of approaches select features corresponding to the criterions that are used in the feature extraction methods. The simplest method so called max variance is to select features by data variance criterion. Later works [21, 28] apply trace ratio LDA criterion to their formulations for feature selection.

Another type of approaches select features by making use of local structure information of data points. Early works [13, 32] empirically show that local structure information is quite helpful for feature selection. They first construct nearest neighbor graph of data points to involve local structure information and then feature selection step follows. Later works [3, 30, 18, 23] also incorporate pre-computed local structure information and conduct feature selection. Recent works [18, 23] select features based on regularized regression with pseudo-label indicators by nonlinear local structure learning methods such as [1, 11].

In this paper, we propose an unsupervised feature selection method so called Simultaneous Orthogonal basis Clustering Feature Selection (SOCFS). SOCFS does not explicitly adopt pre-computed local structure information, but concentrates on the latent cluster information. To this end, a novel target matrix in the regularized regression-based formulation of SOCFS is also proposed to conduct orthogonal basis clustering directly on the projected data points to estimate latent cluster centers. Since the target matrix is put in a single unified term for regression of the proposed objective function, feature selection and clustering are simultaneously performed. In this way, the projection matrix for feature selection is more properly computed by the estimated latent cluster centers of the projected data points. To the best of our knowledge, this formulation is the first

attempt to consider feature selection and clustering together in a single unified term of the objective function. The proposed objective function has fewer parameters to tune and does not require complicated optimization tools so just a simple optimization algorithm is sufficient. Substantial experiments are performed on several publicly available real world datasets, which shows that SOCFS outperforms various unsupervised feature selection methods and that latent cluster information by the target matrix is effective for regularized regression-based feature selection.

2. Preliminary Notations

For a given matrix $\mathbf{W}$, $\mathbf{w}^i$ and $\mathbf{w}_j$ denote i-th row and j-th column of $\mathbf{W}$, respectively, and W_{ij} denotes the (i,j)-th element of $\mathbf{W}$.

For $p > 0$, the l_p-norm of the vector $\mathbf{b} \in \mathbb{R}^n$ is defined as $\|\mathbf{b}\|_p = (\sum_{i=1}^n |b_i|^p)^{\frac{1}{p}}$, and the l_0-norm of the vector $\mathbf{b}$ ($\|\mathbf{b}\|_0$) is defined as the number of non-zero elements in $\mathbf{b}$. The $l_{p,q}$-norm of matrix $\mathbf{W} \in \mathbb{R}^{n\times m}$ is defined as $\|\mathbf{W}\|_{p,q} = (\sum_{i=1}^n \|\mathbf{w}^i\|_p^q)^{\frac{1}{q}}$, where $p > 0, q > 0$. The Frobenius norm of the matrix $\mathbf{W}$ is defined as $\|\mathbf{W}\|_F = \|\mathbf{W}\|_{2,2}$. The $l_{2,0}$-norm of the matrix is defined as $\|\mathbf{W}\|_{2,0} = \sum_{i=1}^n \|\|\mathbf{w}^i\|_2\|_0$.

$|\mathbf{W}|$ denotes a matrix whose elements are the absolute values of the corresponding elements of $\mathbf{W}$. $\mathbf{W} \geq \mathbf{0}$ denotes that all elements of $\mathbf{W}$ are larger than or equal to 0.

3. Problem Formulation

Given training data, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ denote the data matrix with n instances where dimension is d and $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_n] \in \mathbb{R}^{m\times n}$ denote the corresponding target matrix where dimension is m. We start from the regularized regression-based formulation to select maximum r features as follows:

$$\min_{\mathbf{W}} \ \|\mathbf{W}^T\mathbf{X} - \mathbf{T}\|_F^2 \qquad s.t. \ \|\mathbf{W}\|_{2,0} \leq r. \tag{1}$$

Such regularized regression-based feature selection methods have proved to be effective to handle multi-label data in both a supervised and an unsupervised fashion [20, 12, 3, 18, 23]. To exploit such formulation on unlabeled data more effectively, it is crucial for the target matrix $\mathbf{T}$ to have discriminative destinations for projected clusters.

We now propose a new type of target matrix $\mathbf{T}$ that conducts clustering directly on the projected data points $\mathbf{W}^T\mathbf{X}$. To this end, we allow extra degrees of freedom to $\mathbf{T}$ by decomposing it into two other matrices $\mathbf{B} \in \mathbb{R}^{m\times c}$ and $\mathbf{E} \in \mathbb{R}^{n\times c}$ as $\mathbf{T} = \mathbf{B}\mathbf{E}^T$ with additional constraints as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{B},\mathbf{E}} \ & \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \\ s.t. \ & \mathbf{B}^T\mathbf{B} = \mathbf{I}, \ \mathbf{E}^T\mathbf{E} = \mathbf{I}, \ \mathbf{E} \geq \mathbf{0}, \end{aligned} \tag{2}$$

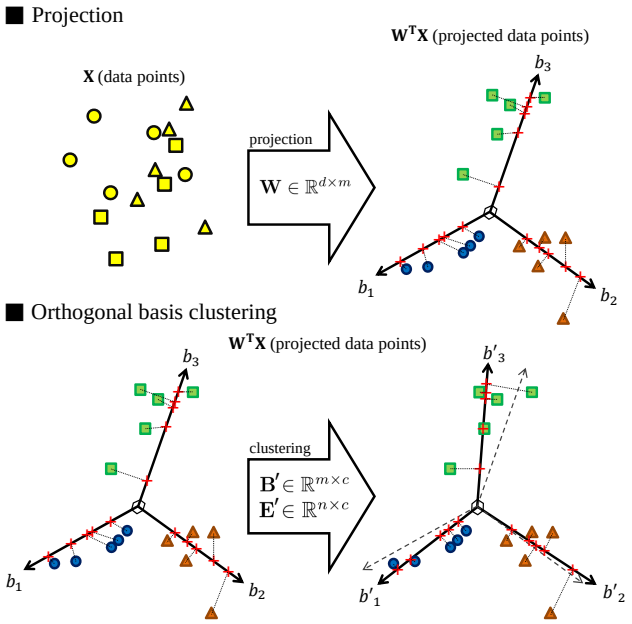


Figure 1. Schematic illustration of the proposed method. First row illustrates the projection step that maps the data points to the target matrix. Second row illustrates the orthogonal basis clustering step to discriminate latent cluster centers of the projected data points. These two steps are simultaneously conducted to select discriminative features without label information.

where $\lambda > 0$ is a weighting parameter for the relaxed regularizer $\|\mathbf{W}\|_{2,1}$ that induces row sparsity of the projection matrix $\mathbf{W}$. If i-th feature is less correlated to the target matrix $\mathbf{T}$, all elements of the row $\mathbf{w}^i$ shrinks to zero. Thus, we can perform feature selection from $\mathbf{W}$ by excluding features corresponding to the zero rows of $\mathbf{W}$.

The meanings of the constraints $\mathbf{B}^T\mathbf{B} = \mathbf{I}, \mathbf{E}^T\mathbf{E} = \mathbf{I}, \mathbf{E} \geq \mathbf{0}$ are as follows: 1) the orthogonal constraint of $\mathbf{B}$ lets each column of $\mathbf{B}$ be independent; 2) the orthogonal and the nonnegative constraint of $\mathbf{E}$ make each row of $\mathbf{E}$ has only one non-zero element [4]. From 1) and 2), we can clearly interpret $\mathbf{B}$ as the basis matrix, which has orthogonality and $\mathbf{E}$ as the encoding matrix, where the non-zero element of each column of $\mathbf{E}^T$ selects one column in $\mathbf{B}$.

While optimizing problem (2), $\mathbf{T} = \mathbf{B}\mathbf{E}^T$ acts like clustering of projected data points $\mathbf{W}^T\mathbf{X}$ with orthogonal basis $\mathbf{B}$ and encoder $\mathbf{E}$, so $\mathbf{T}$ can estimate latent cluster centers of the $\mathbf{W}^T\mathbf{X}$. Then, $\mathbf{W}$ successively projects $\mathbf{X}$ close to corresponding latent cluster centers, which are estimated by $\mathbf{T}$. Note that the orthogonal constraint of $\mathbf{B}$ makes each projected cluster in $\mathbf{W}^T\mathbf{X}$ be separated (independent of each other), and it helps $\mathbf{W}$ to be a better projection matrix for selecting more discriminative features. If the clustering is directly performed on $\mathbf{X}$ not on $\mathbf{W}^T\mathbf{X}$, the orthogonal constraint of $\mathbf{B}$ extremely restricts the degree of freedom of $\mathbf{B}$. However, since features are selected by $\mathbf{W}$ and the clustering is carried out on $\mathbf{W}^T\mathbf{X}$ in our formulation, so the or-

thogonal constraint of $\mathbf{B}$ is highly reasonable. A schematic illustration of the proposed method is shown in Figure 1.

4. Optimization

4.1. Problem Reformulation

Previous methods [5, 31, 17] can solve problem (2) w.r.t $\mathbf{E}$ by approximating the orthogonal constraint. However, because such methods are based on nonnegative matrix factorization (NMF) [16], they focus more on handling the nonnegative constraint than on the orthogonal constraint. Since they cannot fully guarantee $\mathbf{E}$ an orthogonal matrix as also mentioned in their papers, so $\mathbf{E}$ cannot play a role as encoding matrix. Alternatively, we propose an equivalent formulation of problem (2) as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{B},\mathbf{E},\mathbf{F}} \ & \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} \\ s.t. \ & \mathbf{B}^T\mathbf{B} = \mathbf{I}, \ \mathbf{E}^T\mathbf{E} = \mathbf{I}, \ \mathbf{F} = \mathbf{E}, \ \mathbf{F} \geq \mathbf{0}, \end{aligned} \tag{3}$$

where $\mathbf{F}$ is an auxiliary variable with an additional constraint of $\mathbf{F} = \mathbf{E}$. This reformulation step has a goal to detach the nonnegative constraint from $\mathbf{E}$ and assign the constraint to a new variable $\mathbf{F}$. $\mathbf{F}$ has a role to induce nonnegativity to $\mathbf{E}$ while $\mathbf{E}$ keeps orthogonality through the additional constraint $\mathbf{F} = \mathbf{E}$. By rewriting problem (3), we finally propose our objective function of SOCFS as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{B},\mathbf{E},\mathbf{F}} \ & \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1} + \gamma\|\mathbf{F} - \mathbf{E}\|_F^2 \\ s.t. \ & \mathbf{B}^T\mathbf{B} = \mathbf{I}, \ \mathbf{E}^T\mathbf{E} = \mathbf{I}, \ \mathbf{F} \geq \mathbf{0}, \end{aligned} \tag{4}$$

where $\gamma > 0$ is another parameter to control the degree of equivalence between $\mathbf{F}$ and $\mathbf{E}$.

This formulation has two advantages. First, each variable in problem (4) has a single constraint and the objective function is convex in each single variable when the other variables are fixed, so it allows a simple iterative optimization algorithm as in the following subsection. Second, since each single constraint on $\mathbf{E}$ and $\mathbf{B}$ will be directly used in the optimization algorithm, $\mathbf{E}$ as well as $\mathbf{B}$ has orthogonality during the optimization steps. Usually, the orthogonality of $\mathbf{E}$ is more important than the nonnegativity of $\mathbf{E}$ to assure $\mathbf{E}$ as an encoding matrix, so this formulation gives another advantage for the performance of the clustering.

4.2. An Iterative Algorithm

W update: $\mathbf{W}$ is minimized while fixing $\mathbf{B}$, $\mathbf{E}$, and $\mathbf{F}$. The subproblem that only relates to $\mathbf{W}$ is

$$\min_{\mathbf{W}} \ \|\mathbf{W}^T\mathbf{X} - \mathbf{B}\mathbf{E}^T\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}. \tag{5}$$

Similar to [20], setting the derivative of $J(\mathbf{W},\mathbf{B},\mathbf{E},\mathbf{F})$ w.r.t $\mathbf{W}$ to zero, we have

$$\begin{aligned} & \frac{\partial J}{\partial \mathbf{W}} = 2\mathbf{X}\mathbf{X}^T\mathbf{W} - 2\mathbf{X}\mathbf{E}\mathbf{B}^T + 2\lambda\mathbf{D}\mathbf{W} = \mathbf{0} \\ & \Longrightarrow \ \mathbf{W} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{D})^{-1}\mathbf{X}\mathbf{E}\mathbf{B}^T, \end{aligned} \tag{6}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements $D_{ii} = \frac{1}{2\|\mathbf{w}^i\|_2}$. Note that the derivative of $\|\mathbf{W}\|_{2,1}$ w.r.t $\mathbf{W}$ is computed to $2\mathbf{D}\mathbf{W}$.

B update: $\mathbf{B}$ is minimized while fixing $\mathbf{E}$, $\mathbf{W}$, and $\mathbf{F}$. The subproblem that only relates to $\mathbf{B}$ is

$$\min_{\mathbf{B}} \ \|\mathbf{E}\mathbf{B}^T - \mathbf{X}^T\mathbf{W}\|_F^2 \qquad s.t. \ \mathbf{B}^T\mathbf{B} = \mathbf{I}. \tag{7}$$

Proposition 1. *Suppose we have two matrices $\mathbf{P} \in \mathbb{R}^{n\times m}$ and $\mathbf{Q} \in \mathbb{R}^{n\times d}$. The optimization problem*

$$\min_{\tilde{\mathbf{T}}} \ \|\mathbf{P}\tilde{\mathbf{T}} - \mathbf{Q}\|_F^2 \qquad s.t. \ \tilde{\mathbf{T}}\tilde{\mathbf{T}}^T = \mathbf{I} \tag{8}$$

has an analytic solution

$$\tilde{\mathbf{T}} = \mathbf{U}\mathbf{I}_{m,d}\mathbf{V}^T, \tag{9}$$

where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$ are left and right eigenvectors of $\mathbf{P}^T\mathbf{Q}$ computed by SVD, respectively.

Proof. A proof can be done as in [27]. Note that a difference between the problem in Proposition 1 and Orthogonal Procrustes Problem (OPP) [26, 22, 9] is the constraint. □

The solution is obtained by Proposition 1 with $\mathbf{P} = \mathbf{E}$, $\tilde{\mathbf{T}} = \mathbf{B}^T$, and $\mathbf{Q} = \mathbf{X}^T\mathbf{W}$ as

$$\mathbf{B} = \mathbf{V}_\mathbf{B}\mathbf{I}_{m,c}\mathbf{U}_\mathbf{B}{}^T, \tag{10}$$

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

E, F update: $\mathbf{E}$ and $\mathbf{F}$ are minimized while fixing $\mathbf{B}$ and $\mathbf{W}$. $\mathbf{E}$ and $\mathbf{F}$ are successively updated in another iteration fixing each other. The subproblem that only relates to $\mathbf{E}$ is

$$\begin{aligned} \min_{\mathbf{E}} \ & \|\mathbf{B}\mathbf{E}^T - \mathbf{W}^T\mathbf{X}\|_F^2 + \gamma\|\mathbf{E} - \mathbf{F}\|_F^2 \\ s.t. \ & \mathbf{E}^T\mathbf{E} = \mathbf{I}. \end{aligned} \tag{11}$$

The subproblem (11) is rewritten as

$$\begin{aligned} & \arg\min_{\mathbf{E}:\mathbf{E}^T\mathbf{E}=\mathbf{I}} \ \mathrm{tr}(\mathbf{E}\mathbf{B}^T\mathbf{B}\mathbf{E}^T - 2\mathbf{X}^T\mathbf{W}\mathbf{B}\mathbf{E}^T + \mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}) \\ & \qquad + \gamma\,\mathrm{tr}(\mathbf{E}^T\mathbf{E} - 2\mathbf{E}^T\mathbf{F} + \mathbf{F}^T\mathbf{F}) \\ = & \arg\min_{\mathbf{E}:\mathbf{E}^T\mathbf{E}=\mathbf{I}} \ -\mathrm{tr}((\mathbf{X}^T\mathbf{W}\mathbf{B} + \gamma\mathbf{F})\mathbf{E}^T) \\ = & \arg\min_{\mathbf{E}:\mathbf{E}^T\mathbf{E}=\mathbf{I}} \ \|\mathbf{E} - (\mathbf{X}^T\mathbf{W}\mathbf{B} + \gamma\mathbf{F})\|_F^2. \end{aligned} \tag{12}$$

Then the solution of this subproblem is obtained by Proposition 1 with $\mathbf{P} = \mathbf{I}$, $\tilde{\mathbf{T}} = \mathbf{E}^T$, and $\mathbf{Q} = \mathbf{B}^T\mathbf{W}^T\mathbf{X} + \gamma\mathbf{F}^T$ as

Algorithm 1: $\mathbf{E}$, $\mathbf{F}$ update 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 (13) 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 (15);
- $s = s + 1$;

until $\|\triangle J^{(t)}_{EF}(\mathbf{E}'_s, \mathbf{F}'_s)\| \leq \epsilon$ *or* $s \geq S$;

Output: $\mathbf{E}_{t+1} = \mathbf{E}'_s$, $\mathbf{F}_{t+1} = \mathbf{F}'_s$

$$\mathbf{E} = \mathbf{V_E}\mathbf{I}_{n,c}\mathbf{U_E}^T, \tag{13}$$

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. The subproblem that only relates to $\mathbf{F}$ is

$$\min_{\mathbf{F}} \ \|\mathbf{F} - \mathbf{E}\|_F^2 \qquad s.t. \ \mathbf{F} \geq \mathbf{0}. \tag{14}$$

The solution of the subproblem is easily obtained as

$$\mathbf{F} = \frac{1}{2}(\mathbf{E} + |\mathbf{E}|). \tag{15}$$

We summarize the $\mathbf{E}$ and $\mathbf{F}$ update rules of the proposed optimization algorithm in Algorithm 1. The overall proposed optimization algorithm is also presented in Algorithm 2. Following update rules (6), (10), (13), and (15), the objective function monotonically decreases.

4.3. Convergence Analysis

We put the convergence analysis of the proposed optimization algorithm of SOCFS with all update rules in the supplementary material. It is found that SOCFS converges within 100 iterations. We use a single convergence criterion in terms of the number of iterations by setting a maximum iteration value for all experiments in this paper.

Table 1. Dataset Description

	# of Classes	# of Instances	# of Features
LUNG	5	203	3312
COIL20	20	1440	1024
Isolet1	26	1560	617
USPS	10	9258	256
YaleB	38	2414	1024
UMIST	20	575	644
AT&T	40	400	644

5. Experiments

In this section, we evaluate the performance of the proposed method SOCFS. We follow the same experimental setups of the previous works [13, 3, 30, 18, 23].

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 (6);
- Update $\mathbf{B}_{t+1} = \mathbf{V_B}\mathbf{I}_{m,c}\mathbf{U_B}^T$ by (10) 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 $\frac{1}{2\|\mathbf{w}^i_{t+1}\|_2}$;
- $t = t + 1$;

until $\|\triangle J(\mathbf{W}_t, \mathbf{B}_t, \mathbf{E}_t, \mathbf{F}_t)\| \leq \epsilon$ *or* $t \geq T$;

Output: Features are selected corresponding to the largest values of $\|\mathbf{w}^i_t\|, i = 1 \ldots d$, which are sorted by descending order.

5.1. Experimental Setup

The experiments were conducted on seven publicly available datasets. These datasets include one cancer dataset (LUNG[1] [2]), one object image dataset (COIL20[2] [19]), one spoken letter recognition dataset (Isolet1[2] [7]), one handwritten digit dataset (USPS[2] [14]), and three face image datasets (YaleB[2] [8], UMIST[3] [10], and AT&T[4] [25]). Detailed information of the datasets is summarized in Table 1. On each dataset, SOCFS is compared to the following six unsupervised feature selection methods and all features case:

- **Max Variance (MV)** selects features corresponding to the largest variances.
- **Laplacian Score (LS)**[5] **[13]** selects features corresponding to the largest laplacian scores that are computed to reflect the locality preserving power.
- **Multi-Cluster Feature Selection (MCFS)**[5] **[3]** selects features by a two-step method of spectral regression with a l_1 regularizer.
- **Unsupervised Discriminative Feature Selection (UDFS)**[6] **[30]** selects features from local discriminative score that reflects local structure information with a $l_{2,1}$ regularizer.
- **Nonnegative Discriminative Feature Selection (NDFS)**[6] **[18]** selects features by a joint framework of nonnegative spectral analysis and $l_{2,1}$ regularized regression.

[1] https://sites.google.com/site/feipingnie/publications
[2] http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html
[3] http://www.sheffield.ac.uk/eee/research/iel/research/face
[4] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[5] http://www.cad.zju.edu.cn/home/dengcai/Data/MCFS.html
[6] http://www.cs.cmu.edu/ yiyang/Publications.html

- **Robust Unsupervised Feature Selection (RUFS)**[7] **[23]** selects features by a joint framework of $l_{2,1}$ norm-based nonnegative matrix factorization with local learning and $l_{2,1}$ regularized regression.

According to the experimental setups in the previous works, clustering accuracy (ACC) and normalized mutual information (NMI) [29] are measured to evaluate the clustering results of each feature set selected from various methods. For LS, MCFS, UDFS, NDFS, and RUFS, we fix the number of neighboring parameter $k = 5$, following the previous works. Since SOCFS does not consider the local structure information of data points, we do not need to set any neighboring parameters for SOCFS. We set the dimension of projected space m as the number of latent clusters c.

To fairly compare several unsupervised feature selection methods, we tune all parameters for each method by a grid-search strategy from $\{10^{-6}, 10^{-4}, \ldots, 10^{4}, 10^{6}\}$. The number of selected features are set as $\{50, 100, 150, \ldots, 300\}$ for the first six datasets. For USPS dataset, we set the number of features as $\{50, 80, 110, \ldots, 200\}$ due to the total number of features. Clustering experiments are conducted on each selected feature set from several datasets and methods by K-means. To report the best clustering results of each method, different parameters can be used for several datasets and methods. Since the results of K-means depend on initialization of clustering seeds, we repeat the experiments 20 times with random initialization of seeds over all experiments and the mean and standard deviation of the measures ACC and NMI are reported. We also repeat the experiments 20 times for random initialization of variables of the methods such as NDFS, RUFS and SOCFS. We can clearly notice that the methods of better performance will have larger mean and smaller standard deviation values of the measures ACC and NMI of the clustering results. The reliability of the performance of the methods in practical unlabeled data can be also verified by this experimental setup. All the results in the figures and the tables are produced by their published source codes.

5.2. Experimental Results and Analysis

The experimental results are shown in Figures 2-3, and Tables 2-3. We notice that feature selection can effectively improve the clustering results for all datasets while reducing the redundant features. Thus, it is desirable to use selected features for learning tasks.

We evaluate the clustering results of selected features versus number of selected features in Figures 2-3. We notice that, even with a small number of features particularly less than 100 features, SOCFS selects the most discriminative features among the seven methods. It is also shown in Tables 2-3 that SOCFS has the best clustering results for all datasets, which indicates SOCFS selects the most discriminative features under multi-label condition.

We have the following further observations from the figures and the tables. First, the methods MCFS, UDFS, NDFS, RUFS, and SOCFS, which select features simultaneously achieve better results than the others that select features one by one. Since the projection matrices of those methods are determined at the same time during iterations, corresponding features are selected to prevent high correlation. Second, regularized regression-based methods MCFS, NDFS, RUFS, and SOCFS, show relatively better results. For MCFS, NDFS, and RUFS, nonlinear local structure learning methods are used, which can help regression fit data more accurately. Although SOCFS does not use local structure information explicitly, SOCFS estimates latent cluster centers by orthogonal basis clustering of the target matrix, so regression can show better clustering results.

5.3. Sensitiveness of Parameters

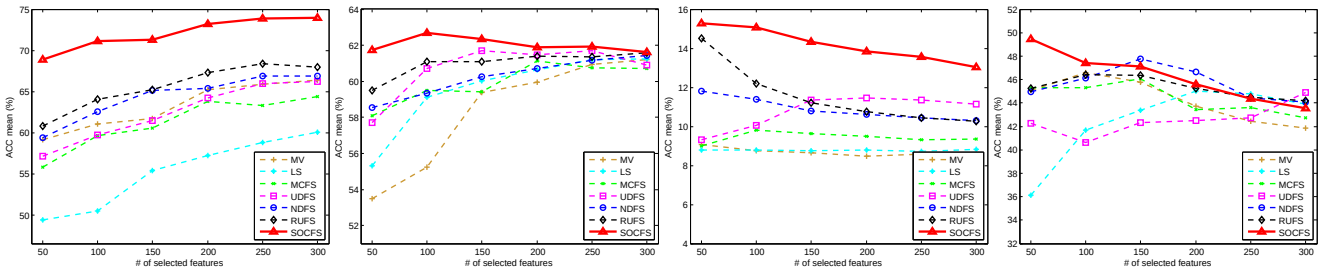
To study the sensitiveness of parameters, clustering results under varying parameters and the number of selected features are measured. Since we cannot tune the parameters according to the measures ACC and NMI w.r.t ground-truth labels on practical unlabeled data, the sensitiveness of parameters is a critical issue. Figures 4-5 tell us that the clustering results of SOCFS are 1) not highly sensitive to λ and γ within wide ranges and 2) slightly more sensitive to λ than to γ. This is because λ controls the sparsity of $\mathbf{W}$ and the proper value of λ is data dependent. Since γ is a relatively insensitive parameter, it takes less time to tune γ.

Therefore, SOCFS is not affected too much by varying the values of the parameters and this means that most λ and γ values can show satisfactory performances. In particular, we found that almost equally prominent performances are obtained by setting λ from 1 to 100. Furthermore, we can fix $\gamma = \lambda$ to simplify the tuning process for efficiency in practice. This is empirically proved in the experiments and also shown in Figures 4-5.

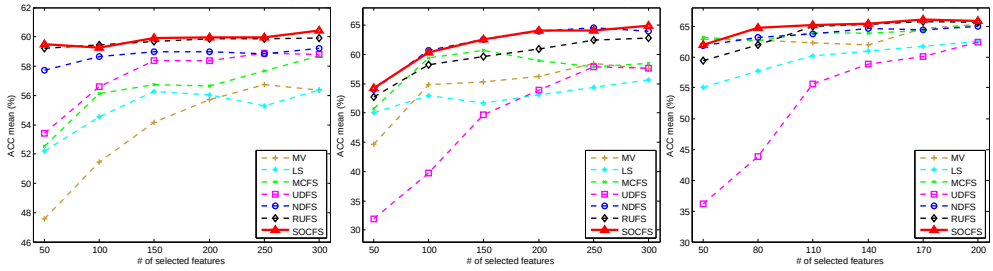
5.4. Discussions

We suggest discussions about the advantages of SOCFS compared to the most recent works RUFS and NDFS. SOCFS is different in the following respects. First, SOCFS has fewer parameters to tune than NDFS and RUFS have. Furthermore, by setting $\gamma = \lambda$ SOCFS has only one parameter λ yet yields better performance, so SOCFS is more appropriate in practical use. Second, SOCFS has a mixed-sign target matrix, which is suitable for the practical mixed-sign data compared to nonnegative target matrices (pseudo-label indicators) of NDFS and RUFS. With those methods, if mixed-sign data is provided, the projected data points should be adjusted to a nonnegative region, so this reduces the possibility of accurate projection and selected features

[7] https://sites.google.com/site/qianmingjie/home/publications

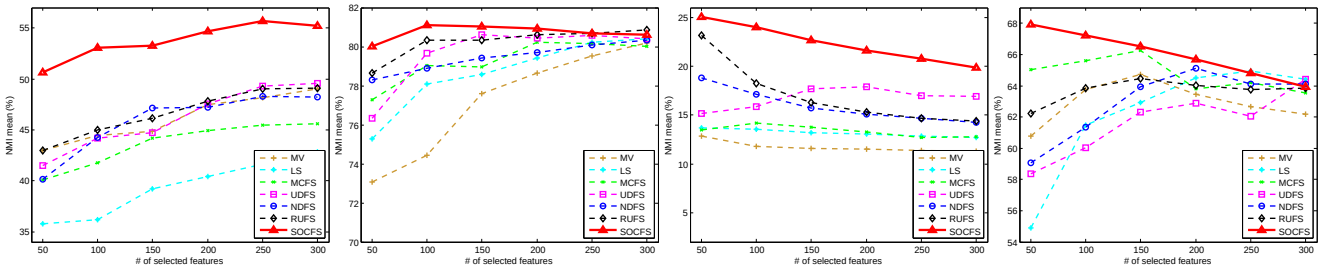


(a) LUNG (b) AT&T (c) YaleB (d) UMIST

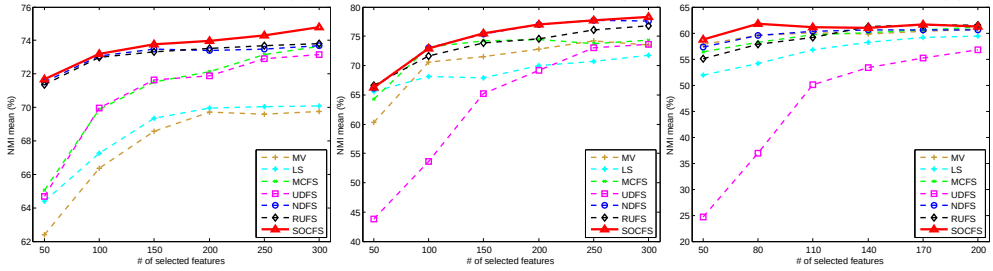


(e) COIL20 (f) Isolet1 (g) USPS

Figure 2. The clustering results ACC mean (%) of each selected feature set from various unsupervised feature selection methods versus # of selected features. SOCFS selects the most discriminative features even with a small number of features.



(a) LUNG (b) AT&T (c) YaleB (d) UMIST



(e) COIL20 (f) Isolet1 (g) USPS

Figure 3. The clustering results NMI mean (%) of each selected feature set from various unsupervised feature selection methods versus # of selected features. SOCFS selects the most discriminative features even with a small number of features.

are less discriminative. This indicates that the nonnegative constraint of the target matrix is too strict to cover the mixed-sign data. But SOCFS has a mixed-sign target matrix for the purpose of estimating latent cluster centers, so it has more flexibility. Third, since both NDFS and RUFS also take advantage of nonnegative matrix factorization (NMF) formulation, their approximation of the orthogonal constraint ($\|\mathbf{F}^T\mathbf{F} - \mathbf{I}\|_F^2$) cannot guarantee the full orthogonality, so their target matrices cannot act as label indicator matrices as we mentioned above. But SOCFS always guarantees the orthogonality of the encoding matrix $\mathbf{E}$ by the constraint, so the target matrix $\mathbf{T}$ effectively determines the latent cluster centers of the projected data points.

6. Conclusion

We have proposed a new unsupervised feature selection method with simultaneous feature selection and clustering combined in a single term of the objective function. In the objective function of this new formulation, a novel type of

Table 2. ACC (% ± std) of various unsupervised feature selection methods on several datasets. The best results are in boldface.

	LUNG	AT&T	YaleB	UMIST	COIL20	Isolet1	USPS
All Features	70.0 ± 8.9	60.9 ± 3.4	9.6 ± 0.6	42.1 ± 2.3	59.4 ± 4.9	57.9 ± 3.6	65.7 ± 2.4
MV	66.4 ± 8.9	61.1 ± 3.4	9.1 ± 0 4	46.7 ± 2.8	56.7 ± 4.6	58.5 ± 3.3	65.3 ± 4.4
LS [13]	60.1 ± 9.5	61.3 ± 3.5	8.8 ± 0.4	45.1 ± 3.4	56.3 ± 4.8	55.6 ± 3.3	62.5 ± 4.4
MCFS [3]	64.3 ± 7.9	61.2 ± 3.7	9.8 ± 0.7	45.1 ± 3.2	58.7 ± 5.3	60.7 ± 4.0	65.2 ± 4.2
UDFS [30]	66.2 ± 7.8	61.7 ± 3.8	11.5 ± 0.7	44.9 ± 2.7	58.9 ± 5.1	57.9 ± 3.0	62.4 ± 3.1
NDFS [18]	66.9 ± 9.1	61.4 ± 3.5	11.8 ± 0.6	47.8 ± 3.1	59.2 ± 5.0	64.6 ± 4.4	64.9 ± 3.1
RUFS [23]	68.4 ± 8.3	61.6 ± 3.2	14.5 ± 0.9	46.4 ± 3.0	59.9 ± 4.9	62.8 ± 3.8	65.8 ± 3.1
SOCFS	**74.0 ± 8.9**	**62.7 ± 3.1**	**15.3 ± 0.7**	**49.4 ± 3.2**	**60.4 ± 4.7**	**64.9 ± 4.4**	**66.1 ± 2.0**

Table 3. NMI (% ± std) of various unsupervised feature selection methods on several datasets. The best results are in boldface.

	LUNG	AT&T	YaleB	UMIST	COIL20	Isolet1	USPS
All Features	51.7 ± 5.4	80.5 ± 1.8	13.0 ± 0.8	63.7 ± 2.5	74.2 ± 2.4	74.2 ± 1.7	60.9 ± 0.8
MV	49.0 ± 5.3	80.2 ± 1.7	12.8 ± 0.5	64.4 ± 2.2	69.8 ± 2.1	74.2 ± 1.4	60.8 ± 1.8
LS [13]	42.9 ± 5.0	80.4 ± 1.8	13.7 ± 0.4	65.1 ± 1.9	70.1 ± 2.3	71.7 ± 1.4	59.5 ± 2.1
MCFS [3]	45.6 ± 4.5	80.2 ± 1.8	14.2 ± 1.0	67.3 ± 2.6	73.7 ± 2.5	74.4 ± 1.9	61.2 ± 1.8
UDFS [30]	49.6 ± 5.1	80.6 ± 1.8	17.9 ± 0.9	64.4 ± 1.4	73.2 ± 2.5	73.6 ± 1.6	56.8 ± 1.4
NDFS [18]	48.3 ± 5.2	80.3 ± 1.8	18.8 ± 0.7	65.1 ± 2.0	73.7 ± 2.1	77.7 ± 2.0	60.7 ± 1.3
RUFS [23]	49.1 ± 5.1	80.9 ± 1.7	23.1 ± 0.7	64.5 ± 2.2	73.8 ± 2.4	76.8 ± 1.9	61.5 ± 1.4
SOCFS	**55.7 ± 6.2**	**81.1 ± 1.6**	**25.1 ± 0.8**	**67.9 ± 2.1**	**74.8 ± 2.3**	**78.3 ± 1.9**	**61.6 ± 0.8**

target matrix has also been proposed to serve as a latent cluster centers by performing orthogonal basis clustering on the projected data points. The orthogonal basis clustering gives latent projected cluster information to yield more accurate selection of discriminative features. The formulation has been turned out to be minimized by proposed simple optimization without other complex optimization algorithms. The effectiveness of the proposed method based on the formulation has carefully proved by the extensive experiments on several real world datasets.

Acknowledgments: This work was supported in part by the Technology Innovation Program, 10045252, Development of robot task intelligence technology, funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2014-003140) and (MSIP) (2010-0028680).

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001. 1

[2] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *National Academy of Sciences (NAS)*, 98(24):13790–13795, 2001. 4

[3] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *KDD*, pages 333–342, 2010. 1, 2, 4, 7

[4] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610, 2005. 2

[5] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *KDD*, pages 126–135, 2006. 3

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. 2nd edition, 2001. 1

[7] M. A. Fanty and R. Cole. Spoken letter recognition. In *NIPS*, page 220, 1990. 4

[8] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):643–660, 2001. 4

[9] G. H. Golub and C. F. Van Loan. Matrix computations. 1996. 3

[10] D. B. Graham and N. M. Allinson. Characterising virtual eigensignatures for general purpose face recognition. In *Face Recognition*, pages 446–456. 1998. 4

[11] Q. Gu and J. Zhou. Local learning regularized nonnegative matrix factorization. In *IJCAI*, 2009. 1

[12] R. He, T. Tan, L. Wang, and W.-S. Zheng. $l_{2,1}$ regularized correntropy for robust feature selection. In *CVPR*, pages 2504–2511, 2012. 2

[13] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*, pages 507–514, 2005. 1, 4, 7

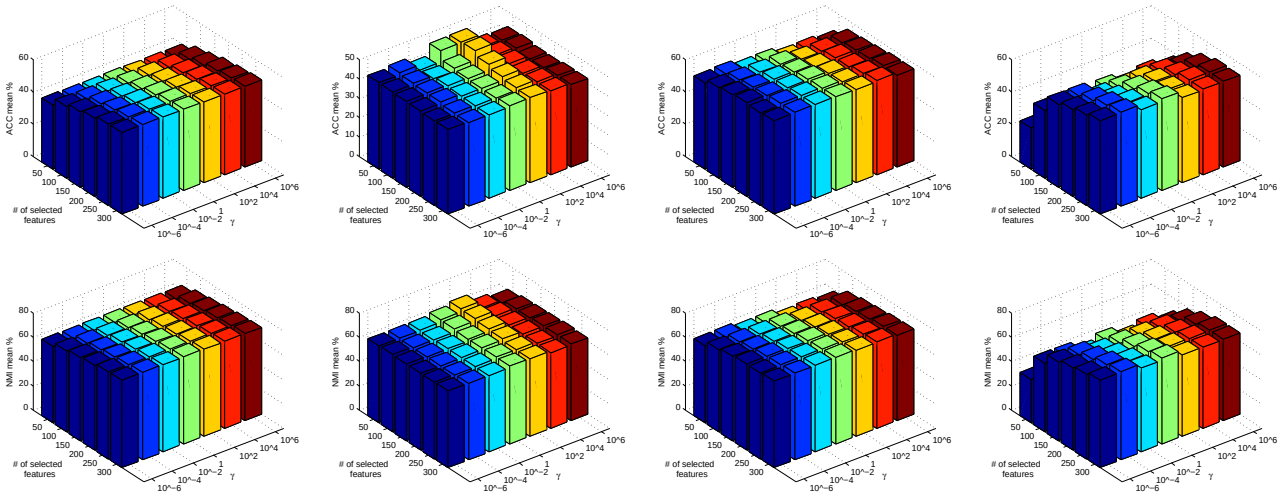
[14] J. J. Hull. A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(5):550–554, 1994. 4

[15] K. Kira and L. A. Rendell. A practical approach to feature selection. In *International Workshop on Machine Learning (ML)*, pages 249–256, 1992. 1

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001. 3

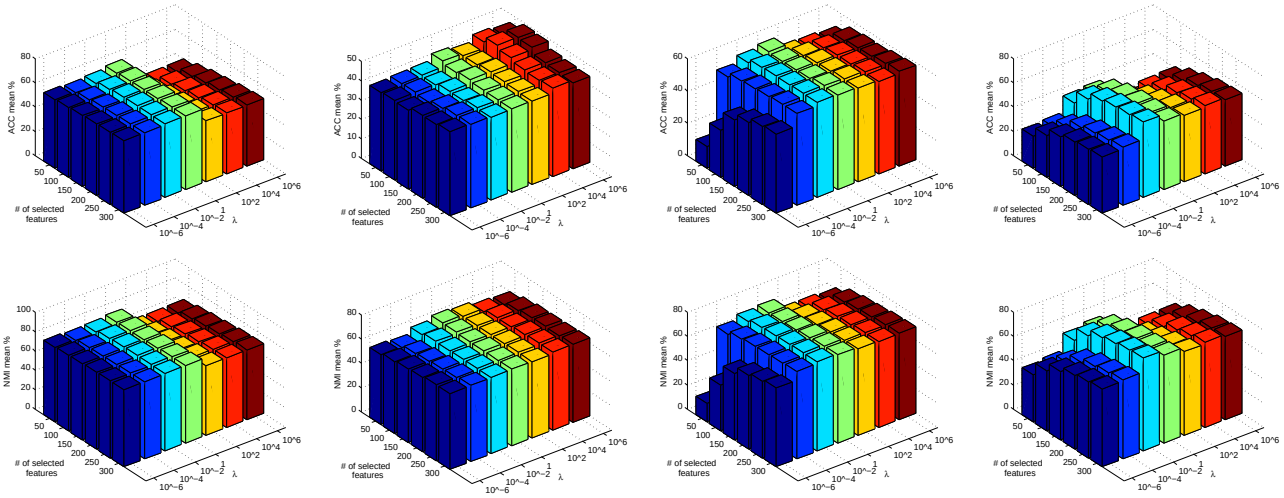
[17] Z. Li, X. Wu, and H. Peng. Nonnegative matrix factorization on orthogonal subspace. *Pattern Recognition Letters*, 31(9):905–911, 2010. 3

[18] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, pages 1026–1032, 2012. 1, 2, 4, 7



(a) AT&T (b) UMIST (c) COIL20 (d) Isolet1

Figure 4. ACC and NMI mean (%) of SOCFS with different γ and number of selected features while keeping $\lambda = 10^2$.



(a) AT&T (b) UMIST (c) COIL20 (d) Isolet1

Figure 5. ACC and NMI mean (%) of SOCFS with different λ and number of selected features while keeping $\gamma = 10^{-2}$.

[19] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical report, CUCS-005-96, 1996. 4

[20] 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. 1, 2, 3

[21] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan. Trace ratio criterion for feature selection. In *AAAI*, pages 671–676, 2008. 1

[22] H. Park. A parallel algorithm for the unbalanced orthogonal procrustes problem. *Parallel Computing*, 17(8):913–923, 1991. 3

[23] M. Qian and C. Zhai. Robust unsupervised feature selection. In *IJCAI*, pages 1621–1627, 2013. 1, 2, 4, 5, 7

[24] L. E. Raileanu and K. Stoffel. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1):77–93, 2004. 1

[25] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994. 4

[26] P. H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. 3

[27] T. Viklands. *Algorithms for the weighted orthogonal Procrustes problem and other least squares problems*. PhD thesis, Umeå University, 2006. 3

[28] D. Wang, F. Nie, and H. Huang. Unsupervised feature selection via unified trace ratio formulation and k-means clustering (track). In *ECML PKDD*, pages 306–321. 2014. 1

[29] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003. 5

[30] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou. $l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011. 1, 4, 7

[31] J. Yoo and S. Choi. Orthogonal nonnegative matrix factorization: Multiplicative updates on stiefel manifolds. In *Intelligent Data Engineering and Automated Learning*, pages 140–147. 2008. 3

[32] Z. Zhao and H. Liu. Spectral feature selection for supervised and unsupervised learning. In *ICML*, pages 1151–1157, 2007. 1