Subspace Clustering via Variance Regularized Ridge Regression

Chong Peng, Zhao Kang, Qiang Cheng

Southern Illinois University, Carbondale, IL, 62901, USA

{pchong,zhao.kang,qcheng}@siu.edu

Abstract

Spectral clustering based subspace clustering methods have emerged recently. When the inputs are 2-dimensional (2D) data, most existing clustering methods convert such data to vectors as preprocessing, which severely damages spatial information of the data. In this paper, we propose a novel subspace clustering method for 2D data with enhanced capability of retaining spatial information for clustering. It seeks two projection matrices and simultaneously constructs a linear representation of the projected data, such that the sought projections help construct the most expressive representation with the most variational information. We regularize our method based on covariance matrices directly obtained from 2D data, which have much smaller size and are more computationally amiable. Moreover, to exploit nonlinear structures of the data, a nonlinear version is proposed, which constructs an adaptive manifold according to updated projections. The learning processes of projections, representation, and manifold thus mutually enhance each other, leading to a powerful data representation. Efficient optimization procedures are proposed, which generate non-increasing objective value sequence with theoretical convergence guarantee. Extensive experimental results confirm the effectiveness of proposed method.

1. Introduction

Representing and processing high-dimensional data has been routinely used in many areas such as computer vision and machine learning. Often times, high-dimensonal data have latent low-dimensional structures and can be well represented by a union of low-dimensional subspaces. Recovering such low-dimensional subspaces usually requires clustering data points into different groups such that each group can be fitted with a subspace, which is refered to as subspace clustering. During the last decade, subspace clustering algorithms have attracted substantial reserach attention, among which spectral-clustering based subspace clustering methods have been popular due to their promising performance; e.g., low-rank representation (LRR) [17] and sparse subspace clustering (SSC) [6] are two typical such methods that seek representation matrices with different assumptions, with LRR assuming low-rankness while SSC requiring sparsity.

Recently, a number of new subspace clustering methods have been developed. For example, [29] replaces the nuclear norm used in LRR by some non-convex rank approximations, because the nuclear norm is far from being accurate in estimating the rank of real world data. [30] reveals that not all features are equally important to recover low-dimensional subspaces and with feature selection both nuclear norm and non-convex rank approximations may obtain enhanced performance. [24] seeks a linear projection to project the data and learns a sparse representation in the projected latent low-dimensional space. To capture nonlinear structures of the data, nonlinear techniques such as kernel and manifold methods have been adopted for subspace clustering. For example, kernel SSC (KSSC) [25] maps data points into a higher-dimensional kernel space where sparse coefficients are learned; [19, 34] construct low-rank representations by exploiting nonlinear structures of the data in a kernel space or on manifold. A shared drawback of these nonlinear methods is that the kernel matrix or graph Laplacian is predefined, and thus may be independent from the representation learning, potentially leading to clustering results far from optimal. A recently developed method, thresholding ridge regression (TRR) [31], points out that such methods as LRR and SSC achieve robustness by estimating and removing specifically structured representation errors from the input space, which requires prior knowledge on the usually unknown structures of the (also unknown) errors. To overcome this limitation, [31] leverages an observation that the representation coefficients are larger over intra-subspace than inter-subspace data points, and thus the representation errors can be eliminated in the projection space by thresholding small coefficients obtained with a ridge regression model.

Subspace clustering has various applications in computer vision areas based on 2-dimensional (2D) data. Because all above-mentioned methods use only vectors as input examples, to make the input samples of 2D matrices suit to these methods, a standard approach is to vectorize the 2D examples. While being commonly employed, this approach does not consider the inherent structure and spatial correlations of the 2D data; more importantly, building models using vectorized data will significantly increase the dimension of the search space of the model, which is not effective to filter the noise, occlusions or redundant information [10]. Besides the way of vectorizing 2D data, tensor based approaches have been proposed. While they may potentially better exploit spatial structures of the 2D data [9, 38], such approaches still have some limitations: They use all features of the data, hence noisy or redundant features may degrade the learning performance. Also, tensor computation and methods usually involve flattening and folding operations, which, more or less, have issues similar to those of vectorization operation and thus might not fully exploit the true structures of the data. Moreover, tensor methods usually suffer from the following major issues: 1) for candecomp/parafac (CP) decomposition based methods, it is generally NP-hard to compute the CP rank [12, 20]; 2) Tucker decomposition is not unique [12]; 3) the application of a core tensor and a high-order tensor product would incur information loss of spatial details [14].

To overcome the limitations of existing methods, we propose a new subspace clustering method for 2D data with enhanced capability of retaining spatial information, which, in particular, has stark differences from tensor-based methods. We summarize the key contributions of this paper as follows: **1)** Projections are sought to simultaneously retain the most variational information from 2D data and help construct the most expressive representation coefficients, such that the learning of projection and representation mutually enhance each other; **2)** Nonlinear relationship between examples are accounted for, where the graph Laplacian is adaptively constructed according to the updated projections. Hence all learning tasks mutually enhance and lead to a powerful data representation; **3)** Covariance matrices constructed from 2D data are used for regularization, which enable the curse of dimensionality to be effectively mitigated; **4)** Efficient optimization procedures are developed with theoretical convergence guarantee of the objective value sequence; 5) Extensive experimental results have verified the effectiveness of the proposed method.

2. Related Work

In this section, we review some methods that are closely related to our work.

2.1. LRR and TRR

Given n examples, the existing subspace clustering methods generally represent each example by a d-dimensional vector and stack all vectors as columns to construct a data matrix $A \in \mathcal{R}^{d\times n}$. LRR seeks the lowest-rank representation of the data with a model,

$$\min_{Z} \|Z\|_* + \tau\|E\|_{2,1} \quad s.t. \quad A = AZ + E, \tag{1}$$

where $\|E\|_{2,1}$ sums ℓ_2 norms of columns of E, and $\|Z\|_*$ is the nuclear norm of Z that sums all its singular values.

It has been pointed out that (1) needs prior knowledge on the structures of the errors, which usually is unknown in practice [29]. TRR overcomes this limitation by eliminating the effects of errors from the projection space with a model of thresholding ridge regression [31],

$$\min_{Z} \|Z\|_F^2 + \tau\|A - AZ\|_F^2 \quad s.t. \quad Z_{ii} = 0, \tag{2}$$

where small values in Z will be truncated to zero by thresholding.

2.2. Two-Dimensional PCA (2DPCA)

Given an image $X \in \mathcal{R}^{a\times b}$ and a unitary vector $p \in \mathcal{R}^b$, the projected feature vector of X can be obtained by the transformation of $y = Xp$ [36, 37]. By defining $G_t = \mathbf{E}((X - \mathbf{E}X)^T(X - \mathbf{E}X))$ with $\mathbf{E}$ being the expectation operator, p can be obtained by

$$\max_{p^Tp=1} \mathrm{Tr}(\mathbf{S}_x) \Leftrightarrow \max_{p^Tp=1} \mathrm{Tr}(p^TG_tp). \tag{3}$$

Usually, finding only one optimal projection direction is not enough [36] and it is necessary to find multiple projection directions $P = [p_1, p_2, \cdots, p_r] \in \mathcal{R}^{b\times r}$. Mathematically, P can be found by solving

$$\max_{P^TP=I_r} \mathrm{Tr}(P^TG_tP), \tag{4}$$

where I_r is an identity matrix of size $r \times r$.

3. 2D Variance Regularized Ridge Regression

In this section, we introduce our new model and develop an optimization scheme.

3.1. Proposed Model

Let $\mathbf{X} = \{X_i \in \mathcal{R}^{a\times b}\}_{i=1}^n$ be a collection of 2D examples (or points), and $P = [p_1, \cdots, p_r] \in \mathcal{R}^{b\times r}$, $Q = [q_1, \cdots, q_r] \in \mathcal{R}^{a\times r}$, $r \leq \min\{a, b\}$, be two projection matrices that satisfy $P^TP = Q^TQ = I_r$. We define $\mathbf{v}(\cdot)$ to be a vectorization operator, $X_i \odot P := \mathbf{v}(X_iP)$, $X_i \otimes Q := \mathbf{v}(X_i^TQ)$, and

$$\begin{aligned} \mathbf{X} \odot P &:= [\mathbf{v}(X_1P), \cdots, \mathbf{v}(X_nP)] \in \mathcal{R}^{ar\times n}, \\ \mathbf{X} \otimes Q &:= [\mathbf{v}(X_1^TQ), \cdots, \mathbf{v}(X_n^TQ)] \in \mathcal{R}^{br\times n}. \end{aligned} \tag{5}$$

Then a new data matrix $Y = [y_1, \cdots, y_n]$ is obtained by defining y_i as

$$y_i = [(X_i \odot P)^T, (X_i \otimes Q)^T]^T \in \mathcal{R}^{(ar+br)}. \quad (6)$$

It is noted that, with projections P and Q, both horizontal and vertical spatial information is retained in Y. We assume that the projected data points have a self-expressive property, i.e., $Y \approx YZ$, where $Z \in \mathcal{R}^{n\times n}$ is the representation matrix, with z_i, $z_{(j)}$, and z_{ij} being its i-th row, j-th column, and ij-th element, respectively. In this paper, we adopt the following ridge regression model [22, 31],

$$\min_Z \|Y - YZ\|_F^2 + \tau\|Z\|_F^2, \quad (7)$$

where $\tau > 0$ is a balancing parameter. Here, unlike [22,31], we do not require $z_{ii} = 0$ because: 1) y_i is in the intra-subspace of y_i itself, thus $z_{ii} \neq 0$ is meaningful; 2) $\tau > 0$ already excludes potentially trivial solutions such as I_n; 3) its effectiveness and efficiency are verified by experiments in Section 6.

It is noted that (7) starkly differs from the existing subspace clustering models, because 2D data are directly used with projections such that inherent spatial information is retained. Now that our method involves two projection matrices P and Q, a necessary question arises: how to find the optimal projection matrices? To answer this question, it is essential to jointly construct the most expressive representation matrix and retain the most information from the original data. Here, we decide to adopt the variation or scattering of the 2D data to represent their information content, inspired by the traditional Fisher's linear discriminant analysis. Mathematically, we jointly minimize the fitting errors of self-expression and maximize total scatters of projected data, which leads to:

$$\min_{Z,P^TP=I_r,Q^TQ=I_r} \frac{\|Y - YZ\|_F^2}{\mathrm{Tr}(P^T\tilde{G}_P P)\mathrm{Tr}(Q^T\tilde{G}_Q Q)} + \tau\|Z\|_F^2, \quad (8)$$

where $\tilde{G}_P$ and $\tilde{G}_Q$[1] are defined to be 2D covariance matrices of X_is and X_j^Ts, respectively. Realizing that $\tilde{G}_P$ and $\tilde{G}_Q$ are usually invertible in real world applications, (8) is equavilent to solving generalized eigenvalue problems [16] with respect to P and Q. To facilitate optimization and increase the flexibility of (8) such that the terms of fitting errors and retaining the most expressive information can be better balanced, we change the quotient in (8) into additive terms and propose the following 2D Variance Regularized Ridge Regression (VR3) model:

$$\begin{aligned}\min_{Z,P^TP=I_r,Q^TQ=I_r} &\|Y - YZ\|_F^2 + \tau\|Z\|_F^2 \\ &+\gamma_1\mathrm{Tr}(P^TG_PP)+\gamma_2\mathrm{Tr}(Q^TG_QQ),\end{aligned} \quad (9)$$

where $G_P = \tilde{G}_P^{-1}$, $G_Q = \tilde{G}_Q^{-1}$[2], and $\gamma_1, \gamma_2 > 0$, are two balancing parameters. It is seen that by minimizing the above objective function, the projections enables us to capture the most variational information from the 2D data and construct the most expressive self-expression, which lead to a powerful data representation.

Remark. Rather than using $\ell_{2,1}$ or ℓ_1 norm to measure the fitting errors, we adopt the Frobenius norm in the first term of (9) for the following reasons: 1) The projections ensure the most variational information is used for constructing the representation, while the adverse effects from less important information, noise, and corruptions are alleviated, thus providing enhanced robustness; 2) [30] shows that the Frobenius norm can well model the fitting errors when the most important features are used for representation learning, which motivates us to use a ℓ_2 norm-based loss model in (9); 3) The Frobenius norm-based model leads to efficient optimization with a mathematically provable convergence guarantee; 4) Extensive experimental results verify the effectiveness of (9) in Section 6.

3.2. Optimization

Now, we develop an efficient alternating optimization procedure to solve (9).

3.2.1 Calculating Q

The subproblem of optimizing Q is

$$\min_{Q^TQ=I_r} \|\mathbf{X} \otimes Q - (\mathbf{X} \otimes Q)Z\|_F^2 + \gamma\mathrm{Tr}(Q^TG_QQ). \quad (10)$$

Theorem 1. *Define $F_1 = \sum_{i=1}^n X_iX_i^T$, $F_2 = \sum_{i=1}^n\sum_{j=1}^n z_{ji}X_iX_j^T$, and $F_3 = \sum_{i=1}^n\sum_{j=1}^n z_{(i)}z_{(j)}^T X_iX_j^T$. The problem of (10) is a constrained quadratic optimization, and admits a closed-form solution,*

$$\boldsymbol{eig}_r(F_1 - 2F_2 + F_3 + \gamma G_Q), \quad (11)$$

where $F_1 - 2F_2 + F_3 + \gamma G_Q$ is positive definite and $\boldsymbol{eig}_r(F)$ returns eigenvectors of F associated with its r smallest eigenvalues.

Proof. It is seen that the first term in (10) is

$$\begin{aligned}&\|\mathbf{X} \otimes Q - (\mathbf{X} \otimes Q)Z\|_F^2 \\ =&\sum_{i=1}^n \Big\|X_i^TQ - \sum_{j=1}^n z_{ji}X_j^TQ\Big\|_F^2 \\ =&\mathbf{Tr}\Big(\sum_{i=1}^n Q^TX_iX_i^TQ\Big) - 2\mathbf{Tr}\Big(Q^TX_i\sum_{j=1}^n z_{ji}X_j^TQ\Big)\end{aligned} \quad (12)$$

[1] In practice, they are estimated by $\tilde{G}_P = \sum_{i=1}^n (X_i - \bar{X})^T(X_i - \bar{X})$ and $\tilde{G}_Q = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

[2] In singular case, in the implementation, we define $G_P = (\tilde{G}_P + \epsilon I_n)^{-1}$, where $\epsilon > 0$ is a small value. Similar approach can be found in [21], where the Schatten norm is smoothed by adding ϵI_n to guarantee differentiability. In the experiments, we always observe that G_P is nonsingular and thus positive definite. Similar strategy is adopted for G_Q.

$$
\begin{aligned}
&+\mathbf{Tr}\Big(\Big(\sum_{j=1}^{n} z_{ji}X_j^TQ\Big)^T\Big(\sum_{j=1}^{n} z_{ji}X_j^TQ\Big)\Big)\\
=&\mathbf{Tr}\Big(Q^T\Big(\sum_{i=1}^{n} X_iX_i^T\Big)Q\Big)-2\mathbf{Tr}\Big(Q^T\Big(\sum_{j=1}^{n} z_{ji}X_iX_j^T\Big)Q\Big)\\
&+\mathbf{Tr}\Big(Q^T\Big(\sum_{i,j=1}^{n} z_{(i)}z_{(j)}^TX_iX_j^T\Big)Q\Big)\\
=&\mathbf{Tr}\Big(Q^TF_1Q\Big)-2\mathbf{Tr}\Big(Q^TF_2Q\Big)+\mathbf{Tr}\Big(Q^TF_3Q\Big)\\
=&\mathbf{Tr}(Q^T(F_1-2F_2+F_3)Q).
\end{aligned}
$$

Therefore, (10) is reduced to

$$
\min_{Q^TQ=I_r} \mathbf{Tr}\big(Q^T\left(F_1-2F_2+F_3+\gamma_2G_Q\right)Q\big). \tag{13}
$$

It is easy to verify that $F_1-2F_2+F_3+\gamma_2G_Q$ is positive definite due to the nonnegativity of (10), and thus the optimal solution is obtained by (11). □

3.2.2 Calculating P

The subproblem of optimizing P is

$$
\min_{P^TP=I_r} \|\mathbf{X}\odot P-(\mathbf{X}\odot P)Z\|_F^2+\gamma_1\mathrm{Tr}(P^TG_PP). \tag{14}
$$

Theorem 2. *Define* $H_1=\sum_{i=1}^n X_i^TX_i$, $H_2=\sum_{i=1}^n\sum_{j=1}^n z_{ji}X_i^TX_j$, *and* $H_3=\sum_{i=1}^n\sum_{j=1}^n z_{(i)}z_{(j)}^TX_i^TX_j$. *The problem of* (14) *is a constrained quadratic optimization and admits a closed-form solution,*

$$
\boldsymbol{eig}_r(H_1-2H_2+H_3+\gamma G_P), \tag{15}
$$

where $H_1-2H_2+H_3+\gamma G_P$ *is positive definite.*

Proof. It can be shown similarly to that of Theorem 1. □

3.2.3 Calculating Z

The subproblem of optimizing Z is

$$
\min_Z \|Y-YZ\|_F^2+\tau\|Z\|_F^2, \tag{16}
$$

which admits the following solution

$$
Z=(Y^TY+\tau I_n)^{-1}(Y^TY). \tag{17}
$$

Theorem 3. *Denoting the objective function of* (9) *by* $\mathcal{J}(Q,P,Z)$, *under the updating rules of* (11), (15) *and* (17), *the value sequence of the objective function* $\{\mathcal{J}(Q^k,P^k,Z^k)\}_{k=1}^{\infty}$ *is non-increasing and converges, where* k *denotes the iteration number.*

Proof. According to Theorems 1 and 2, it is easy to see that

$$
\begin{aligned}
&\mathcal{J}(Q^{k+1},P^{k+1},Z^{k+1})\le\mathcal{J}(Q^{k+1},P^{k+1},Z^k)\\
\le&\mathcal{J}(Q^{k+1},P^k,Z^k)\le\mathcal{J}(Q^k,P^k,Z^k).
\end{aligned} \tag{18}
$$

It is obvious that $\mathcal{J}(Q,P,Z)\ge 0$ by its definition. Therefore, $\{\mathcal{J}(Q^k,P^k,Z^k)\}_{k=1}^{\infty}$ converges. □

4. Nonlinear VR3

In this section, we expand our VR3 model to account for nonlinearity in the instance space.

4.1. Proposed Nonlinear Model

The above proposed VR3 model learns a representation of the projected data in the Euclidean space, which only considers the linear relationship of the data. Because nonlinear relationship in the instance space usually exists and is important in real world applications, it is important to take into consideration such nonlinearity. We decide to account for nonlinear structures of the data on manifold inspired by [3]. We suppose the following assumption on the representation matrix Z is true: If two data points y_i and y_j are close on the manifold, then their new representations given by z_i and z_j should be also close, which leads to minimizing the following quantity:

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\|z_i-z_j\|_2^2\\
=&\sum_{j=1}^{n} d_{jj}z_j^Tz_j-\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}z_i^Tz_j\\
=&\mathrm{Tr}(ZDZ^T)-\mathrm{Tr}(ZWZ^T)=\mathrm{Tr}(ZLZ^T),
\end{aligned} \tag{19}
$$

where $W=[w_{ij}]$ is the similarity matrix with w_{ij} being the similarity between y_i and y_j, and D is a diagonal matrix with $d_{ii}=\sum_j W_{ij}$. Here, we aim to consider the nonlinear relationship of the projected data with both projection matrices; therefore, we construct two manifolds using $\mathbf{X}\odot P$ and $\mathbf{X}\otimes Q$, respectively, which leads to our model of Nonlinear VR3 (NVR3):

$$
\begin{aligned}
\min_{Z,P^TP=I_r,Q^TQ=I_r} &\mathcal{J}(Q,P,Z)+\eta_1\mathrm{Tr}(ZL_PZ^T)\\
&+\eta_2\mathrm{Tr}(ZL_QZ^T),
\end{aligned} \tag{20}
$$

where L_P (resp. L_Q) is obtained from D_P-W_P (resp. D_Q-W_Q), which are based on $\mathbf{X}\odot P$ (resp. $\mathbf{X}\otimes Q$). Various weighting schemes can be used to define the similarities between projected data points [3]; however, it is out of the scope of this paper regarding how to choose the best weighting scheme. In this paper, for simplicity yet without loss of generality, we use dot-product weighting to define W_P by $[W_P]_{ij}=(X_i\odot P)^T(X_j\odot P)$, such that a fully connected graph is constructed. Similar strategy is adopted to construct W_Q. It is seen that the manifolds are learned adaptively, such that the processes of learning projections, representations, and manifold mutually enhance, thus leading to a powerful data representation.

4.2. Optimization

Now, we discuss the optimization procedure of (20).

4.3. Calculating Q

The subproblem for Q-minimization is

$$\min_{Q^TQ=I_r} \|\mathbf{X}\otimes Q-(\mathbf{X}\otimes Q)Z\|_F^2 + \gamma_2\mathrm{Tr}(Q^TG_QQ)+\eta_2\mathrm{Tr}(ZL_QZ^T). \tag{21}$$

Theorem 4. *Define $F_4=\sum_{i=1}^n\sum_{j=1}^n\|z_i-z_j\|_2^2X_iX_j^T$. Given that Z is bounded, the problem in (21) is a constrained quadratic optimization, and admits a closed-form solution,*

$$\boldsymbol{eig}_r(F_1-2F_2+F_3+\gamma_2G_Q+\frac{\eta_2}{2}F_4), \tag{22}$$

where F_1, F_2, and F_3 are defined in Theorem 1.

Proof. Omitting the factor η_2, the last term in (21) is

$$\begin{aligned}&\mathrm{Tr}(ZL_QZ^T)\\ =&\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\|z_i-z_j\|_2^2(X_i\otimes Q)^T(X_j\otimes Q)\\ =&\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\|z_i-z_j\|_2^2\sum_{s=1}^r q_s^TX_iX_j^Tq_s\\ =&\frac{1}{2}\sum_{s=1}^r q_s^T\Big(\sum_{i=1}^n\sum_{j=1}^n\|z_i-z_j\|_2^2X_iX_j^T\Big)q_s\\ =&\frac{1}{2}\mathbf{Tr}(Q^TF_4Q).\end{aligned} \tag{23}$$

Let $\lambda_i(\cdot)$ be the ith largest eigenvalue of the input matrix, then (21) is equivalent to minimizing (21)-$\frac{\eta_1}{2}\lambda_a(F_4)$. Based on the proof of Theorem 1, the subproblem associated with Q is to minimize $\mathbf{Tr}(Q^TCQ)$, with constraint $Q^TQ=I_r$, where $C=F_1-2F_2+F_3+\gamma_2G_Q+\frac{\eta_2}{2}F_4-\frac{\eta_1}{2}\lambda_a(F_4)I_a$. It is easy to verify that C is positive definite and $\mathbf{eig}_r(C)$ gives the optimal solution to (21). It is also easy to verify that $\mathbf{eig}_r(C)$ is equivalent to (22), which concludes the proof. □

4.4. Calculating P

The subproblem of optimizing P is

$$\min_{Q^TQ=I_r} \|\mathbf{X}\odot P-(\mathbf{X}\odot P)Z\|_F^2 + \gamma_1\mathrm{Tr}(P^TG_PP)+\eta_1\mathrm{Tr}(ZL_PZ^T). \tag{24}$$

Theorem 5. *Define $H_4=\sum_{i=1}^n\sum_{j=1}^n\|z_i-z_j\|_2^2X_i^TX_j$. Given that Z is bounded, the problem of (24) is a constrained quadratic optimization, and admits a closed-form solution,*

$$\boldsymbol{eig}_r(H_1-2H_2+H_3+\gamma_1G_P+\frac{\eta_1}{2}H_4), \tag{25}$$

where H_1, H_2, and H_3 are defined in Theorem 2.

Proof. Similar to proof of Theorem 4. □

4.5. Calculating Z

The subproblem of optimizing Z is

$$\min_Z\|Y-YZ\|_F^2+\tau\|Z\|_F^2 + \eta_1\mathrm{Tr}(ZL_PZ^T)+\eta_2\mathrm{Tr}(ZL_QZ^T). \tag{26}$$

Under the condition in Theorem 6, we update Z by setting the derivative of (26) to zero:

$$2Y^TYZ-2Y^TY+2\tau Z+2\eta_1ZL_P+2\eta_2ZL_Q=0. \tag{27}$$

It is seen that (27) is a Sylvester equation and can be solved by the MATLAB built-in function 'lyap':

$$Z=\mathrm{lyap}(Y^TY+\frac{\tau}{2}I_n,\eta_1L_P+\eta_2L_Q+\frac{\tau}{2}I_n,Y^TY). \tag{28}$$

Theorem 6. *If $\tau\geq-\eta_1r(\min_i\{\lambda_b(\sum_{j=1}^nX_i^TX_j)\}-\lambda_1(\sum_{j=1}^nX_j^TX_j))-\eta_2r(\min_i\{\lambda_a(\sum_{j=1}^nX_iX_j^T)\}-\lambda_1(\sum_{j=1}^nX_jX_j^T))$, then (28) is bounded and is the optimal solution to (26).*

Proof. Let $\tau_1=r\min_i\{\lambda_b(\sum_{j=1}^nX_i^TX_j)\}-r\lambda_1(\sum_{j=1}^nX_j^TX_j)$, and $\tau_2=r\min_i\{\lambda_a(\sum_{j=1}^nX_iX_j^T)\}-r\lambda_1(\sum_{j=1}^nX_jX_j^T)$. Because $[W_P]_{ij}=\sum_{s=1}^rp_s^TX_i^TX_jp_s$, $\mathrm{Tr}(W_P)=\sum_{s=1}^rp_s^T(\sum_{j=1}^nX_j^TX_j)p_s\leq r\lambda_1(\sum_{j=1}^nX_j^TX_j)$, implying that $\lambda_i(W_P)\leq\lambda_1(W_P)\leq r\lambda_1(\sum_{j=1}^nX_j^TX_j)$. For D_P, $[D_P]_{ii}=\sum_{j=1}^n[W_P]_{ij}=\sum_{s=1}^rp_s^T(\sum_{j=1}^nX_i^TX_j)p_s\geq r\lambda_b(\sum_{j=1}^nX_i^TX_j)$. Let $S^T(\mathrm{diag}\{\lambda_i(W_P)\})S$ be the eigenvalue decomposition of W_P, then $D_P-W_P=S^T(\mathrm{diag}\{[D_P]_{ii}-\lambda_i(W_P)\})S$, implying that $\lambda_i(D_P-W_P)\geq\min_i\{[D_P]_{ii}\}-\lambda_i(W_P)\geq r\min_i\{\lambda_b(\sum_{j=1}^nX_i^TX_j)\}-r\lambda_1(\sum_{j=1}^nX_j^TX_j)$, i.e., $\lambda_i(L_P)\geq\tau_1$. Similarly, we can prove that $\lambda_i(L_Q)\geq\tau_2$.

Therefore, $-\eta_1\tau_1I_n+\eta_1L_P$ and $-\eta_2\tau_2I_n+\eta_2L_Q$ are positive definite. We decompose the objective in (26) as $\|Y-YZ\|_F^2+(\tau+\eta_1\tau_1+\eta_2\tau_2)\|Z\|_F^2+\eta_1\mathrm{Tr}(ZL_PZ^T)-\eta_1\tau_1\|Z\|_F^2+\eta_2\mathrm{Tr}(ZL_QZ^T)-\eta_2\tau_2\|Z\|_F^2=(\tau+\eta_1\tau_1+\eta_2\tau_2)\|Z\|_F^2+\mathrm{Tr}(Z(-\eta_1\tau_1I_n+\eta_1L_P)Z^T)+\mathrm{Tr}(Z(-\eta_2\tau_2I_n+\eta_2L_P)Z^T)+\|Y-YZ\|_F^2$, where the last three terms on the right hand side of the above equality are convex. For the first term, it is also convex because $\tau+\eta_1\tau_1+\eta_2\tau_2\geq-\eta_1\tau_1-\eta_2\tau_2+\eta_1\tau_1+\eta_2\tau_2=0$. Therefore, the objective in (26) is convex.

Hence, (28) is the optimal solution of (26) by the first order optimality condition. It is easy to verify that at each iteration, the solution Z to (26) is bounded. □

Theorem 7. *Let $\mathcal{L}(Q,P,Z)$ denote the objective function of (26). If the condition of Theorem 6 is satisfied, then under the updating rules of (22), (25) and (28), $\{\mathcal{L}(Q^k,P^k,Z^k)\}_{k=1}^\infty$ is non-increasing and converges, where k denotes the iteration number.*

Proof. Under the condition of Theorem 6, Theorems 4-6 hold, implying that $\mathcal{L}(Q^k, P^k, Z^k)$ is non-increasing. According to Theorem 6, it is easy to verify the boundedness of $\mathcal{L}(Q^k, P^k, Z^k)$. Therefore, $\{\mathcal{L}(Q^k, P^k, Z^k)\}_{k=1}^{\infty}$ converges. □

Remark. It has been recently studied in spectral graph theory [5] and manifold learning theory [2] that a nearest neighbor graph on a scatter of data points can effectively model the local geometric structure. With a more general construction of W_Q and W_P, [27, 33] suggest a viable way for optimization: we first update Q and P by (11) and (15); then we update L_Q and L_P accordingly.

5. Subspace Clustering via VR3 and NVR3

Constructing an affinity matrix from a representation coefficient matrix is commonly applied as a post-processing step for many spectral clustering-based subspace clustering methods [17, 26, 29]. Similarly, after obtaining Z, we construct an affinity matrix $\mathbf{A}$ with the following steps 1) - 3):

1) Let $Z = U\Sigma V^T$ be the skinny SVD of Z. Define the weighted column space of Z as $\bar{Z} = U\Sigma^{1/2}$.

2) Obtain $\bar{U}$ by normalizing the rows of $\bar{Z}$.

3) Construct the affinity matrix $\mathbf{A}$ as $[\mathbf{A}]_{ij} = \left(|[\bar{U}\bar{U}^T]_{ij}|\right)^{\phi}$, where $\phi \geq 1$ controls the sharpness of the affinity matrix between two data points[3].

Subsequently, Normalized Cut (NCut) [32] is performed on $\mathbf{A}$ in a way similar to [1, 29].

6. Experiments

In this section, we conduct experiments to verify the effectiveness of the proposed VR3 and NVR3.

6.1. Algorithms in Comparison

To evaluate the effectiveness of the proposed VR3 and NVR3, several state-of-the-art or recently developed subspace clustering methods are taken as baseline algorithms, including local subspace affinity (LSA) [35], spectral curvature clustering (SCC) [4], LRR [17], low-rank subspace clustering (LRSC) [7], SSC [6], kernel SSC (KSSC) [25], latent LRR (LatLRR) [18], block-diagonal LRR (BDLRR) [8], block-diagonal SSC (BDSSC) [8], structured sparse subspace clustering (S^{3}C) [15], nearest subspace neighbor (NSN) [23], and TRR [31].

We define clustering $accuracy = \frac{1}{n}\sum_{i=1}^{n}\Delta(map(s_i), l_i)$, where l_i and s_i denote the true and predicted labels of X_i, respectively, $map(s_i)$ maps each cluster label s_i to the equivalent label from the data set by permutation such that the accuracy metric can be maximized (because clustering results give clusters up to permutations of labels), and Δ is the Kronecker delta function. More details about this metric can be found in [27, 28]. For fair comparison, the number of clusters, K, is specified for all methods. For the algorithms in comparison, whenever available, we obtain their results from [6, 15, 23, 25], where the parameters have been finely tuned; otherwise, we finely tune their parameters and report the best results. For our method, we iterate the algorithms with a maximum of 200 iterations or when the difference between two consecutive objective values is smaller than 0.001. The parameters are also tuned for our proposed models. Our code is available online[4].

6.2. Face Clustering

Face clustering is an important topic in computer vision. It refers to finding groups from face images, such that each group corresponds to an individual person. In this task, we use Extended Yale B[5] (EYaleB) data set [13] to evaluate the performance of our method. EYaleB data collect face images from 38 persons, of which each has 64 frontal face images taken under varying lighting conditions. These images are cropped to 192×168 pixels. To reduce cost in computation, we down-sample these images to 48×42 pixel as commonly done in the literature [6, 17, 29]. Subsets containing different number of subjects, i.e., $K \in \{2, 3, 5, 8, 10\}$ are collected to better investigate the performance of proposed method. To avoid the potentially combinatorially large number of subsets, we divide these 38 subjects into four groups, containing subjects 1-10, 11-20, 21-30, and 31-38, respectively. Then all possible combinations of subsets with $K \in \{2, 3, 5, 8, 10\}$ are collected within each group, and the collections from the four groups are combined with respect to each K value to obtain five collections of subsets. This data preprocessing is a common way in literature [6, 29]. For TRR, we project the data to be $10K$ in dimension and fix regularization and thresholding parameters to be 100 and 9, respectively. Then within each collection, we conduct experiments on all subsets and report the mean and median clustering accuracies in Table 1.

From Table 1, it is observed that SSC, NSN, S^{3}C and TRR are among the most competitive baseline methods. Competitive performance in both mean and median accuracies can be observed for these methods when K increases from 2 to 10, while the performances of the other baseline methods degrade significantly. Although LatLRR, DBLRR, BDSSC, and LRR-H have good performances with small K values, their performance in the case of large K values may limit their applications to real world problems. VR3 and NVR3 are seen to enhance the clustering performance sig-

[3] For fair comparison, we follow [11, 30] and set $\phi = 4$ in this work.

[4] https://www.researchgate.net/publication/315760668_NVR3_code_pub

[5] http://www.ccs.neu.edu/home/eelhami/codes.htm

Table 1. Clustering Performance on Extended Yale B data set

No. of Subjects	2 Subjects		3 Subjects		5 Subjects		8 Subjects		10 Subjects	
Error Rate (%)	Average	Median	Average	Median	Average	Median	Average	Median	Average	Median
LSA	67.20	52.34	47.71	50.00	41.98	43.13	40.81	41.41	39.58	42.50
SCC	83.38	92.18	61.84	60.94	41.10	40.62	33.89	35.35	26.98	24.22
LRR	90.48	94.53	80.48	85.42	65.84	65.00	58.81	56.25	61.15	58.91
LRR-H	97.46	99.22	95.79	97.40	93.10	94.37	85.66	89.94	77.08	76.41
LRSC	94.68	95.31	91.53	92.19	87.76	88.75	76.28	71.97	69.64	71.25
SSC	98.14	**100.0**	96.90	98.96	95.69	97.50	94.15	95.51	89.06	94.37
LatLRR	97.46	99.22	95.79	97.40	93.10	94.37	85.66	89.94	77.08	76.41
BDLRR	96.09	–	89.98	–	87.03	–	72.30	–	69.16	–
BDSSC	96.10	–	82.30	–	72.50	–	66.80	–	60.47	–
S^3C	98.57	**100.0**	96.91	**99.48**	95.92	97.81	95.16	95.90	93.91	94.84
NSN	98.29	99.22	96.37	96.88	94.19	95.31	91.54	92.38	90.18	90.94
TRR	97.87	99.22	97.07	98.44	96.17	97.50	95.69	96.48	95.10	95.78
VR3	**99.12**	**100.0**	99.23	**99.48**	98.96	**99.38**	98.73	98.63	98.85	98.75
NVR3	99.07	**100.0**	**99.26**	**99.48**	**99.25**	**99.38**	**99.16**	**99.22**	**99.38**	**99.38**

The best performance is bold-faced. "-" means that the result is not reported in the corresponding paper. The parameters for VR3 (resp. NVR3) are (6, 10, 1, 5) (resp. (6,5,1,2,2e-4,5e-4)) ordered as $(r, \tau, \gamma_1, \gamma_2)$ (resp. $(r, \tau, \gamma_1, \gamma_2, \eta_1, \eta_2)$).

nificantly. The mean and median accuracies of our models are the best, and they decrease gracefully when K increases, which shows strong insensitivity to K values. This observation suggests that our method is potentially more suitable than those methods in comparison for real world applications. In general, NVR3 has better performance than VR3 because of its capability of capturing nonlinear structures of the data.

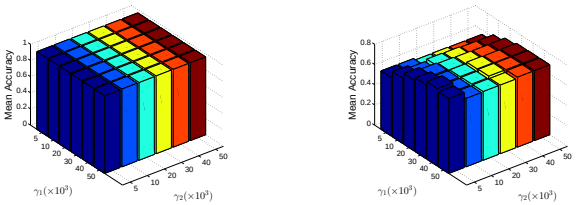
6.3. Handwritten Digit Clustering

We also test the proposed method in handwritten digit clustering of Alphadigits data[6]. This data set contains 36 clusters, including binary digits 0-9 and capital letters A-Z. Each cluster contains 39 images of size 20×16 pixels. Similar to the experimental settings for face clustering, we divide this dataset into 4 groups, containing 0-9, A-J, K-T, and U-Z, respectively. Then subsets with $K \in \{2, 3, 5, 8, 10\}$ are collected in a way similar to EYaleB data. Then we apply all methods in comparison on this dataset, and report the mean and median performances within each collection of subsets in Table 2.

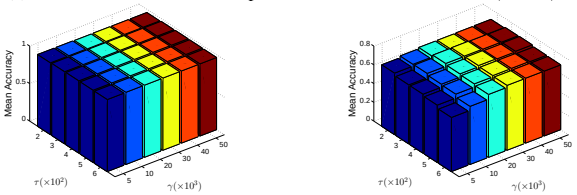
Again, it is observed that VR3 and NVR3 outperform the other methods in all cases, which indicates the importance of 2D approach proposed in this paper. Also, the performance of NVR3 improves that of VR3, demonstrating the importance of learning nonlinear structures of projected data.

For VR3, we use $(r, \tau, \gamma_1, \gamma_2) = (3, 600, 5e4, 5e3)$; for NVR3, $(r, \tau, \gamma_1, \gamma_2) = (3, 600, 6e4, 5e3)$. For LS3C, we project data to be $2K$ in dimension and fix 0.2 and 0.1 for two regularization parameters. For S^3C, we project data to be $2K$ in dimension and fix $\alpha = 0.25$ and $\tau = 1000$. For TRR, we project data to be $8K$ in dimension and fix regularization and thresholding parameters to be 1000 and 13, respectively.

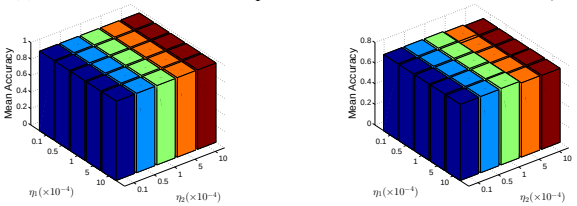
[6] http://www.cs.nyu.edu/~roweis/data.html



(a) Performance of VR3 with respect to different combinations of γ_1 and γ_2.



(b) Performance of VR3 with respect to different combinations of τ and γ.



(c) Performance of NVR3 with respect to different combinations of η_1 and η_2.

Figure 1. Performance of VR3 and NVR3. $K = 2$ on the left while $K = 10$ on the right for (a)-(c).

6.4. Parameter Sensitivity

For unsupervised learning methods, insensitivity to parameter variations is demanding for enhancing their stability in real world applications. Here, we conduct experiments on Alphadigits data set to illustrate the insensitivities of VR3 and NVR3 to parameter variations. We fix $r = 3$ in all cases. In Fig. 1(a), we fix $\tau = 600$ to show the performance of VR3 with respect to different combinations of γ_1 and γ_2. It is evident that competitive performance is obtained over a wide range of parameters. This also suggests we jointly tune γ_1 and γ_2. To demonstrate the effectiveness of this tunning strategy, we set $\gamma_1 = \gamma_2 = \gamma$ and (b) shows the performance of VR3 with various combinations

Table 2. Clustering Performance on Alphadigits data set

No. of Subjects	2 Subjects		3 Subjects		5 Subjects		8 Subjects		10 Subjects	
Error Rate (%)	Average	Median	Average	Median	Average	Median	Average	Median	Average	Median
LSA	89.30	96.15	77.31	77.78	66.19	66.15	59.24	59.94	57.35	58.72
SSC	94.30	97.44	86.42	91.46	76.74	74.88	70.00	69.99	67.86	67.18
LRR	92.24	96.16	85.79	88.89	76.66	76.41	69.50	69.56	66.33	67.44
LRSC	84.19	91.03	74.35	74.36	62.23	62.05	52.02	51.92	49.23	48.97
KSSC (P)	94.58	97.44	87.15	92.31	77.36	76.92	68.94	67.95	66.15	65.64
KSSC (G)	94.07	97.44	86.36	91.45	76.16	73.85	68.81	68.91	67.52	66.67
LS3C	93.77	96.15	87.33	90.17	74.76	73.85	65.94	66.03	62.99	63.51
S^3C	93.34	94.87	86.34	88.03	72.89	71.28	64.17	65.06	62.82	61.79
TRR	95.60	97.44	90.71	93.59	81.02	**83.59**	72.06	72.44	68.38	69.49
VR3	96.10	98.72	91.14	94.02	81.19	83.08	73.13	74.04	73.85	76.67
NVR3	**96.21**	**98.72**	**91.84**	**94.87**	**81.57**	**83.59**	**73.27**	**74.04**	**76.41**	**76.92**

of τ and γ in Fig. 1(b). It is seen that VR3 has promising results that are insensitive to various combinations of τ and γ. Thus, γ_1 and γ_2 can be tuned jointly. For NVR3, similar observations can be also made. In fact, if we set η_1 and η_2 to be zero, then Fig. 1(a)-(b) are special cases of NVR3. In Fig. 1(c), we show the performance of NVR3 with different combinations of η_1 and η_2, with the others the sane as Section 6.3. It is evident that competitive performance is obtained with different combinations and we can jointly tune η_1 and η_2 for NVR3.

Remark. As discussed above, $\{\lambda_1, \lambda_2\}$ and $\{\eta_1, \eta_2\}$ can be jointly tuned in practice. Also, r is usually small as most variational information can be obtained by a few top projection directions. Therefore, we only need to tune 3 and 2 parameters for NVR3 and VR3, respectively. Since this load of tuning is quite common in unsupervised learning [18, 24, 29], the promising performance of the proposed VR3 and NVR3 indicates their potential in real world applications.

6.5. Feature Extraction and Image Reconstruction

In Fig. 2, we show some component and reconstructed images to demonstrate the effects of projections obtained by P and Q visually. Here, we use EYaleB data and VR3 as our examples. We set $r = 30$ to seek orthogonal projection directions both horizontally and vertically. It is observed that the major information is retained by the first several component images such that the original image can be well reconstructed by these components. Also, the principal features captured by P and Q have strong vertical or horizontal patterns, respectively, which verifies the effectiveness of such projections.

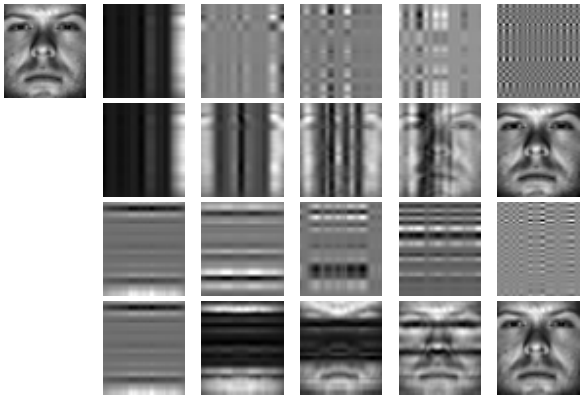


Figure 2. The top left is the original image. For the rest, the first (resp. third) row are the ith column (row) component image $Xp_ip_i^T$ (resp. $q_iq_i^TX$), and the second (resp. fourth) row are the reconstructed images $\sum_{j=1}^{i} Xp_jp_j^T$ (resp. $\sum_{j=1}^{i} q_jq_j^TX$) using the first i column (resp. row) component images, which from left to right represents i = 1, 3, 8, 15, and 30, respectively.

7. Conclusions

In this paper, we present a new subspace clustering method with two models, VR3 and NVR3, with applications to 2D data. Our method is capable of directly using 2D data, and thus the spatial information is maximally retained. Two projection matrices are sought, which simultaneously keep the most variational information from the data and learn the most expressive representation matrix as well as discriminant manifold for the nonlinear variant, such that these individual learning tasks mutually enhance each other and lead to a powerful data representation. Moreover, 2D data are used to construct 2D covariance matrices in our method, which is more computationally amiable than vectorized data. Efficient optimization procedures are developed to solve the proposed models with theoretical guarantee for the convergence of objective value sequence. Extensive experimental results verify that VR3 and NVR3 outperform the state-of-the-art subspace clustering methods. The superior performance with wide insensitivity to model parameters suggests the potential of our method for real world applications.

Acknowledgment

Qiang Cheng is the corresponding author. This work is supported by National Science Foundation under grant IIS-1218712, National Science Foundation of China under grant 11241005, and Foundation Program of Yuncheng University under grants SWSX201603 and YQ-2012020.

References

[1] P. K. Agarwal and N. H. Mustafa. k-means projective clustering. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 155–165. ACM, 2004.

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

[3] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011.

[4] G. Chen and G. Lerman. Spectral curvature clustering (scc). *International Journal of Computer Vision*, 81(3):317–330, 2009.

[5] F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2765–2781, 2013.

[7] P. Favaro, R. Vidal, and A. Ravichandran. A closed form solution to robust subspace estimation and clustering. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1801–1807. IEEE, 2011.

[8] J. Feng, Z. Lin, H. Xu, and S. Yan. Robust subspace segmentation with block-diagonal prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3818–3825, 2014.

[9] Y. Fu, J. Gao, X. Hong, and D. Tien. Low rank representation on riemannian manifold of symmetric positive definite matrices. In *Proceedings of SDM*. SIAM, 2015.

[10] Y. Fu, J. Gao, D. Tien, Z. Lin, and X. Hong. Tensor lrr and sparse coding-based subspace clustering. *IEEE transactions on neural networks and learning systems*, 27(10):2120–2133, 2016.

[11] Z. Kang, C. Peng, and Q. Cheng. Robust subspace clustering via tighter rank approximation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 393–401. ACM, 2015.

[12] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[13] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005.

[14] D. Letexier and S. Bourennane. Noise removal from hyperspectral images by multidimensional filtering. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):2061–2069, 2008.

[15] C.-G. Li and R. Vidal. Structured sparse subspace clustering: A unified optimization framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 277–286, 2015.

[16] M. Li and B. Yuan. 2d-lda: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, 26(5):527–532, 2005.

[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184, 2013.

[18] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1615–1622. IEEE, 2011.

[19] J. Liu, Y. Chen, J. Zhang, and Z. Xu. Enhancing low-rank subspace clustering by manifold regularization. *Image Processing, IEEE Transactions on*, 23(9):4022–4030, 2014.

[20] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5249–5257, 2016.

[21] C. Lu, Z. Lin, and S. Yan. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, 24(2):646–654, 2015.

[22] C.-Y. Lu, H. Min, Z.-Q. Zhao, L. Zhu, D.-S. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *European conference on computer vision*, pages 347–360. Springer, 2012.

[23] D. Park, C. Caramanis, and S. Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems*, pages 2753–2761, 2014.

[24] V. M. Patel, H. Van Nguyen, and R. Vidal. Latent space sparse subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 225–232, 2013.

[25] V. M. Patel and R. Vidal. Kernel sparse subspace clustering. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 2849–2853. IEEE, 2014.

[26] C. Peng, Z. Kang, and Q. Cheng. Integrating feature and graph learning with low-rank representation. *Neurocomputing*, 2017. doi 10.1016/j.neucom.2017.03.071.

[27] C. Peng, Z. Kang, Y. Hu, J. Cheng, and Q. Cheng. Nonnegative matrix factorization with integrated graph and feature learning. *ACM Transactions on Intelligent Systems and Techonology (TIST)*, 8(3):42, 2017.

[28] C. Peng, Z. Kang, Y. Hu, J. Cheng, and Q. Cheng. Robust graph regularized nonnegative matrix factorization for clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(3):33, 2017.

[29] C. Peng, Z. Kang, H. Li, and Q. Cheng. Subspace clustering using log-determinant rank approximation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 925–934. ACM, 2015.

[30] C. Peng, Z. Kang, M. Yang, and Q. Cheng. Feature selection embedded subspace clustering. *IEEE Signal Processing Letters*, 23(7):1018–1022, July 2016.

[31] X. Peng, Z. Yi, and H. Tang. Robust subspace clustering via thresholding ridge regression. In *AAAI*, pages 3827–3833, 2015.

[32] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[33] J. J.-Y. Wang, H. Bensmail, and X. Gao. Feature selection and multi-kernel learning for sparse representation on a manifold. *Neural Networks*, 51:9–16, 2014.

[34] S. Xiao, M. Tan, D. Xu, and Z. Y. Dong. Robust kernel low-rank representation. *IEEE transactions on neural networks and learning systems*, 27(11):2268–2281, 2016.

[35] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Computer Vision–ECCV 2006*, pages 94–106. Springer, 2006.

[36] J. Yang and J.-y. Yang. From image vector to matrix: a straightforward image projection techniqueimpca vs. pca. *Pattern Recognition*, 35(9):1997–1999, 2002.

[37] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 26(1):131–137, 2004.

[38] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao. Low-rank tensor constrained multiview subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1582–1590, 2015.