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-dimensional 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 referred 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 dimensions 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 cande-comp/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 \( A \in \mathbb{R}^{d \times n} \). LRR seeks the lowest-rank representation of the data with a model,

\[
\min_{Z} \|Z\|_* + \tau \|E\|_{2,1} \quad s.t. \quad A = Z + E, (1)
\]

where \( \|E\|_{2,1} \) sums \( \ell_2 \) 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].

\[
\min_{Z} \|Z\|_F^2 + \tau \|A - Z\|_F^2 \quad s.t. \quad Z_{ii} = 0, (2)
\]

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

2.2. Two-Dimensional PCA (2DPCA)

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

\[
\max_{p^T} \text{Tr}(S_x) \leftrightarrow \max_{p^T} \text{Tr}(p^T G_t p). \quad (3)
\]

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

\[
\max_{p^T} \sum_{i \in I_r} \text{Tr}(P^T G_t P), \quad (4)
\]

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 \( X = \{X_i \in \mathbb{R}^{a \times b}\}_{i=1}^n \) be a collection of 2D examples (or points), and \( P = [p_1, \ldots, p_r] \in \mathbb{R}^{b \times r} \), \( Q = [q_1, \ldots, q_r] \in \mathbb{R}^{a \times r} \), \( r \leq \min \{a, b\} \), be two projection matrices that satisfy \( P^T P = Q^T Q = I_r \). We define \( \psi(\cdot) \) to be a vectorization operator, \( X_i \odot P := \psi(X_i P), X_i \odot Q := \psi(X_i^T Q), \) and

\[
X \odot P := [\psi(X_1 P), \ldots, \psi(X_n P)] \in \mathbb{R}^{ar \times n},
\]

\[
X \odot Q := [\psi(X_1^T Q), \ldots, \psi(X_n^T Q)] \in \mathbb{R}^{br \times n}. \quad (5)
\]
Then a new data matrix \( Y = [y_1, \cdots, y_t] \) is obtained by defining \( y_t \) as
\[
y_t = [(X_t \odot P)^T, (X_t \odot Q)^T]^T \in \mathbb{R}^{(ar+br)}.
\]
(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 \( Z \in \mathbb{R}^{n \times n} \) is the representation matrix, with \( z_i, z_{ij} \), 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],
\[
\min_Z \|Y - YZ\|_F^2 + \tau \|Z\|_F^2,
\]
(7)
where \( \tau > 0 \) is a balancing parameter. Here, unlike [22, 31], we do not require \( z_{ii} = 0 \) because: 1) \( y_t \) is in the intra-subspace of \( y_t \), 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:
\[
\min_{Z, P^T P = I, Q^T Q = I} \|Y - YZ\|_F^2 + \tau \|Z\|_F^2 + \lambda \|P\|_F^2 + \gamma_1 \text{Tr}(P^T G_P P) + \gamma_2 \text{Tr}(Q^T G_Q Q),
\]
(8)
where \( \tilde{G}_P \) and \( \tilde{G}_Q \) are defined to be 2D covariance matrices of \( X_i \)'s and \( X_j \)'s, respectively. Realizing that \( \tilde{G}_P \) and \( \tilde{G}_Q \) are usually invertible in real world applications, (8) is equivalent 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:
\[
\min_{Z, P^T P = I, Q^T Q = I} \|Y - YZ\|_F^2 + \tau \|Z\|_F^2
\]
\[
+ \gamma_1 \text{Tr}(P^T G_P P) + \gamma_2 \text{Tr}(Q^T G_Q Q),
\]
where \( G_P = \tilde{G}_P^{-1}, G_Q = \tilde{G}_Q^{-1} \), and \( \gamma_1, \gamma_2 > 0 \), are two balancing parameters. It is seen that by minimizing the above objective function, the projections enable 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 \( \ell_1 \) 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 an \( \ell_2 \) 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
\[
\min_{Q^T Q = I} \|X \odot Q - (X \odot Q)Z\|_F^2 + \gamma \text{Tr}(Q^T G_Q Q).
\]
(10)

Theorem 1. Define \( F_1 = \sum_{i=1}^n X_i X_i^T, F_2 = \sum_{i=1}^n \sum_{j=1}^n z_{ij} X_i X_j^T \), and \( F_3 = \sum_{i=1}^n \sum_{j=1}^n (z_{ij})^2 X_i X_j^T \). The problem of (10) is a constrained quadratic optimization, and admits a closed-form solution,
\[
eig r(F_1 - 2F_2 + F_3 + \gamma G_Q),
\]
(11)
where \( F_1 - 2F_2 + F_3 + \gamma G_Q \) is positive definite and \( \eig r(F) \) returns eigenvectors of \( F \) associated with its \( r \) smallest eigenvalues.

Proof. It is seen that the first term in (10) is
\[
\|X \odot Q - (X \odot Q)Z\|_F^2
\]
\[
= \sum_{i=1}^n \left\|X_i^T Q - \sum_{j=1}^n z_{ij} X_j^T Q\right\|_F^2
\]
\[
= \text{Tr} \left( \sum_{i=1}^n X_i^T Q \left( \sum_{j=1}^n z_{ij} X_j^T Q \right)\right) - 2 \text{Tr} \left( Q^T X_i \sum_{j=1}^n z_{ij} X_j^T Q \right)
\]
\[
\]
\[2^n \text{In singular case, in the implementation, we define } G_P = (\tilde{G}_P + \epsilon I_n)^{-1}, \text{ where } \epsilon > 0 \text{ is a small value. Similar approach can be found in [21], where the Schatten norm is smoothed by adding } \epsilon I_n \text{ to guarantee differentiability. In the experiments, we always observe that } G_P \text{ is nonsingular and thus positive definite. Similar strategy is adopted for } G_Q.\]
and

\begin{equation*}
\text{Proof.}
\end{equation*}

\text{It can be shown similarly to that of Theorem 1.}

\subsection{Calculating P}

The subproblem of optimizing P is

\begin{equation}
\min_{P^T P = I, \gamma} \|X \odot P - (X \odot P)Z\|_F^2 + \gamma \text{Tr}(P^T G P P). \tag{14}
\end{equation}

\textbf{Theorem 2.} Define \( H_1 = \sum_{i=1}^n X_i^T X_i \), \( H_2 = \sum_{j=1}^m z_j^T X_j^T X_j \), and \( H_3 = \sum_{i=1}^n \sum_{j=1}^m z(i)_j z(j)_j X_i^T X_j \). The problem of (14) is a constrained quadratic optimization and admits a closed-form solution,

\begin{equation}
\text{eig.}(H_1 - 2H_2 + H_3 + \gamma G_P), \tag{15}
\end{equation}

\text{where} \( H_1 - 2H_2 + H_3 + \gamma G_P \) \text{is positive definite.}

\textbf{Proof.} It can be shown similarly to that of Theorem 1. \square

\subsection{Calculating Z}

The subproblem of optimizing Z is

\begin{equation}
\min_{z} \|Y - YZ\|_F^2 + \tau \|Z\|_F^2, \tag{16}
\end{equation}

which admits the following solution

\begin{equation}
Z = (Y^T Y + \tau I_n)^{-1} (Y^T Y). \tag{17}
\end{equation}

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

\textbf{Proof.} According to Theorems 1 and 2, it is easy to see that

\begin{equation}
\begin{align*}
J(Q^{k+1}, P^{k+1}, Z^{k+1}) & \leq J(Q^{k+1}, Z^{k+1}) \leq J(Q^k, Z^k) \\
& \leq J(Q^k, P^k, Z^k).
\end{align*} \tag{18}
\end{equation}

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

\subsection{Nonlinear VR3}

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

\textbf{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{equation}
\begin{align}
&\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m w_{ij} \|z_i - z_j\|_2^2 \\
= &\sum_{j=1}^m d_{jj} z_j^T z_j - \sum_{i=1}^n \sum_{j=1}^m w_{ij} z_i^T z_j \\
= &\text{Tr}(Z D Z^T) - \text{Tr}(Z W Z^T) = \text{Tr}(Z L Z^T),
\end{align} \tag{19}
\end{equation}

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 \( X \odot P \) and \( X \odot Q \), respectively, which leads to our model of Nonlinear VR3 (NVR3):

\begin{equation}
\begin{align}
\min_{Z, P^T P = I, Q^T Q = I, \tau} J(Q, P, Z) + \nu_1 \text{Tr}(Z L P Z^T) \\
+ \nu_2 \text{Tr}(Z L_Q Z^T),
\end{align} \tag{20}
\end{equation}

where \( L_P \) (resp. \( L_Q \)) is obtained from \( D_P - W_P \) (resp. \( D_Q - W_Q \)), which are based on \( X \odot P \) (resp. \( X \odot Q \)). 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.

\subsection{Optimization}

Now, we discuss the optimization procedure of (20).
4.3. Calculating $Q$

The subproblem for $Q$-minimization is

\[
\min_{Q:Q=I_r} \|X \otimes Q - (X \otimes Q)Z\|_F^2 + \gamma_2 \text{Tr}(Q^T G_Q Q) + \eta_2 \text{Tr}(ZL_Q Z^T). \tag{21}
\]

**Theorem 4.** Define $F_1 = \sum_{j=1}^n \sum_{j=1}^n \|z_i - z_j\|_2^2 X_i X_j^T$. Given that $Z$ is bounded, the problem in (21) is a constrained quadratic optimization, and admits a closed-form solution,

\[
eig_r(F_1 - 2F_2 + F_3 + \gamma_2 G_Q + \frac{\eta_2}{2} F_4), \tag{22}
\]

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

**Proof.** Omitting the factor $\eta_2$, the last term in (21) is

\[
\text{Tr}(ZL_Q Z^T) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \|z_i - z_j\|_2^2 (X_i \otimes Q)^T (X_j \otimes Q)
\]

\[
= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \|z_i - z_j\|_2^2 \sum_{s=1}^r q_s^T X_i X_j^T q_s
\]

\[
= \frac{1}{2} \sum_{s=1}^r q_s^T \left( \sum_{i=1}^n \sum_{j=1}^n \|z_i - z_j\|_2^2 X_i X_j^T \right) q_s
\]

\[
= \frac{1}{2} \text{Tr}(Q^T F_4 Q).
\]

Let $\lambda_i(\cdot)$ be the $i$th largest eigenvalue of the input matrix, then (21) is equivalent to minimizing \(\frac{\gamma_2}{2} \lambda_r(F_4)\). Based on the proof of Theorem 1, the subproblem associated with $Q$ is to minimize $\text{Tr}(Q^T C Q)$, with constraint $Q^T Q = I_r$, where $C = F_1 - 2F_2 + F_3 + \gamma_2 G_Q + \frac{\eta_2}{2} F_4 - \lambda_1(F_4) I_n$. It is easy to verify that $C$ is positive definite and $\eig_r(C)$ gives the optimal solution to (21). It is also easy to verify that $\eig_r(C)$ is equivalent to (22), which concludes the proof.

\[\square\]

4.4. Calculating $P$

The subproblem of optimizing $P$ is

\[
\min_{P:Q=I_r} \|X \otimes P - (X \otimes P)Z\|_F^2 + \eta_1 \text{Tr}(Q^T G_P P) + \eta_1 \text{Tr}(ZL_P Z^T). \tag{24}
\]

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

\[
eig_r(H_1 - 2H_2 + H_3 + \gamma_1 G_P + \frac{\eta_1}{2} H_4), \tag{25}
\]

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

**Proof.** Similar to proof of Theorem 4.

\[\square\]

4.5. Calculating $Z$

The subproblem of optimizing $Z$ is

\[
\min_Z \|Y - Y Z\|_F^2 + \tau \|Z\|_F^2 + \eta_1 \text{Tr}(ZL_P Z^T) + \eta_2 \text{Tr}(ZL_Q Z^T). \tag{26}
\]

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

\[2Y^T Y Z - 2Y^T Z + 2\tau Z + 2\eta_1 ZL_P + 2\eta_2 ZL_Q = 0. \tag{27}
\]

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

\[
Z = \text{lyap}(Y^T Y + \frac{\tau}{2} I_n, \eta_1 L_P + \eta_2 L_Q + \frac{\tau}{2} I_n, Y Z Y^T). \tag{28}
\]

**Theorem 6.** If \(\tau \geq -\eta_1 r(\min_i \{\lambda_i(\sum_{j=1}^n X_i^T X_j)\} - \lambda_1(\sum_{j=1}^n X_i^T X_j))\) and \(\eta_2 r(\min_i \{\lambda_i(\sum_{j=1}^n X_i^T X_j)\} - \lambda_1(\sum_{j=1}^n X_i^T X_j))\), then (28) is bounded and is the optimal solution to (26).

**Proof.** Let $\tau_1 = r \min_i \{\lambda_i(\sum_{j=1}^n X_i^T X_j)\} - \lambda_1(\sum_{j=1}^n X_i^T X_j)$, and $\tau_2 = r \min_i \{\lambda_i(\sum_{j=1}^n X_i^T X_j)\} - \lambda_1(\sum_{j=1}^n X_i^T X_j)$. Because $[W_P]_{ij} = \sum_{s=1}^r p_s^T X_i^T X_j p_s$, $\text{Tr}(W_P) = \sum_{s=1}^r p_s^T (\sum_{i=1}^n X_i^T X_i) p_s \leq r \lambda_1(\sum_{j=1}^n X_i^T X_j)$, implying that $\lambda_1(W_P) \leq \lambda_1(W_P) \leq r \lambda_1(\sum_{j=1}^n X_i^T X_j)$. For $D_P$, $[D_P]_{ij} = \sum_{s=1}^r p_s^T (\sum_{j=1}^n X_i^T X_j) p_s \geq r \lambda_0(\sum_{j=1}^n X_i^T X_j)$. Let $S^T(\text{diag}(\lambda_i(W_P))) S$ be the eigenvalue decomposition of $W_P$, then $D_P - W_P = S^T(\text{diag}(||D_P||_2 - \lambda_1(W_P))) S$, implying that $\lambda_1(D_P - W_P) \geq \min_i \{||D_P||_2 - \lambda_1(W_P)\} \geq r \min_i \{\lambda_i(\sum_{j=1}^n X_i^T X_j)\} - \lambda_1(\sum_{j=1}^n X_i^T X_j)$, i.e., $\lambda_1(L_P) \geq \tau_1$. Similarly, we can prove that $\lambda_1(L_Q) \geq \tau_2$. Therefore, $-\eta_1 \tau_1 l_n + \eta_1 L_P$ and $-\eta_2 \tau_2 l_n + \eta_2 L_Q$ are positive definite. We decompose the objective in (26) as $\|Y - Y Z\|_F^2 + (\tau + \eta_1 \tau_1 + \eta_2 \tau_2) ||Z||_F^2 + \eta_1 \text{Tr}(ZL_P Z^T) - \eta_1 \tau_1 ||Z||_F^2 + \eta_2 \text{Tr}(ZL_Q Z^T) - \eta_2 \tau_2 ||Z||_F^2 = (\tau + \eta_1 \tau_1 + \eta_2 \tau_2) ||Z||_F^2 + \text{Tr}(Z(-\eta_1 \tau_1 I_n + \eta_1 L_P) Z^T) + \text{Tr}(Z(-\eta_2 \tau_2 I_n + \eta_2 L_Q) Z^T) + \|Y - Y Z\|_F^2$, 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.

\[\square\]

**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 \( A \) 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 \( A \) as \( [A]_{ij} = \phi \), where \( \phi \geq 1 \) controls the sharpness of the affinity matrix between two data points.

Subsequently, Normalized Cut (NCut) [32] is performed on \( A \) 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 (S3C) [15], nearest subspace neighbor (NSN) [23], and TRR [31].

We define clustering accuracy \( \frac{1}{n} \sum_{i=1}^{n} \Delta (\text{map}(s_i), l_i) \), where \( l_i \) and \( s_i \) denote the true and predicted labels of \( X_i \), respectively, \( \text{map}(s_i) \) 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 \( \Delta \) 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.

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 (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 the 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 10\( K \) 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, S3C 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-

---

\(^4\)https://www.researchgate.net/publication/31576668_NVR3_code_pub

\(^3\)http://www.ccs.neu.edu/home/eelhami/codes.htm
Table 1. Clustering Performance on Extended Yale B data set

<table>
<thead>
<tr>
<th>No. of Subjects</th>
<th>Error Rate (%)</th>
<th>Average</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Subjects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSA</td>
<td>67.20</td>
<td>52.54</td>
<td>47.71</td>
</tr>
<tr>
<td>SCC</td>
<td>83.38</td>
<td>92.18</td>
<td>61.84</td>
</tr>
<tr>
<td>LRR</td>
<td>90.48</td>
<td>94.53</td>
<td>80.48</td>
</tr>
<tr>
<td>LRR-H</td>
<td>97.46</td>
<td>99.22</td>
<td>95.79</td>
</tr>
<tr>
<td>LRSC</td>
<td>94.68</td>
<td>95.31</td>
<td>91.53</td>
</tr>
<tr>
<td>SSC</td>
<td>98.14</td>
<td>100.0</td>
<td>96.90</td>
</tr>
<tr>
<td>LatLRR</td>
<td>97.46</td>
<td>99.22</td>
<td>95.79</td>
</tr>
<tr>
<td>BDLRR</td>
<td>96.09</td>
<td>–</td>
<td>89.98</td>
</tr>
<tr>
<td>BDSSC</td>
<td>96.10</td>
<td>–</td>
<td>82.30</td>
</tr>
<tr>
<td>S3C</td>
<td>98.57</td>
<td>100.0</td>
<td>96.91</td>
</tr>
<tr>
<td>TRR</td>
<td>97.87</td>
<td>99.22</td>
<td>97.07</td>
</tr>
<tr>
<td><strong>VR3</strong></td>
<td><strong>99.12</strong></td>
<td><strong>100.0</strong></td>
<td><strong>99.23</strong></td>
</tr>
<tr>
<td><strong>NVR3</strong></td>
<td><strong>99.07</strong></td>
<td><strong>100.0</strong></td>
<td><strong>99.26</strong></td>
</tr>
</tbody>
</table>

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, τ, γ1, γ2, η1, η2).

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 data⁶. 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 ∈ {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, τ, γ1, γ2) = (3, 600, 5e4, 5e3); for NVR3, (r, τ, γ1, γ2) = (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 S3C, we project data to be 2K in dimension and fix α = 0.25 and τ = 1000. For TRR, we project data to be 8K in dimension and fix regularization and thresholding parameters to be 1000 and 13, respectively.

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

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 τ = 600 to show the performance of VR3 with respect to different combinations of γ1 and γ2. It is evident that competitive performance is obtained over a wide range of parameters. This also suggests we jointly tune γ1 and γ2. To demonstrate the effectiveness of this tuning strategy, we set γ1 = γ2 = γ and (b) shows the performance of VR3 with various combinations.
of \( \tau \) and \( \gamma \) in Fig. 1(b). It is seen that VR3 has promising results that are insensitive to various combinations of \( \tau \) and \( \gamma \). Thus, \( \gamma_1 \) and \( \gamma_2 \) can be tuned jointly. For NVR3, similar observations can be also made. In fact, if we set \( \eta_1 \) and \( \eta_2 \) 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 \( \eta_1 \) and \( \eta_2 \), with the others the same as Section 6.3. It is evident that competitive performance is obtained with different combinations and we can jointly tune \( \eta_1 \) and \( \eta_2 \) 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.

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


