Robust Manifold Matrix Factorization for Joint Clustering and Feature Extraction

Lefei Zhang
School of Computer
Wuhan University, Wuhan, China
zhanglefei@whu.edu.cn

Qian Zhang
Alibaba Group
Beijing, China
qianzhang.zq@alibaba-inc.com

Bo Du *
School of Computer
Wuhan University, Wuhan, China
remoteking@whu.edu.cn

Dacheng Tao
Centre for Artificial Intelligence
University of Technology Sydney, Australia
dacheng.tao@uts.edu.au

Jane You
Department of Computing
The Hong Kong Polytechnic University, Hong Kong
csyjia@comp.polyu.edu.hk

Abstract

Low-rank matrix approximation has been widely used for data subspace clustering and feature representation in many computer vision and pattern recognition applications. However, in order to enhance the discriminability, most of the matrix approximation based feature extraction algorithms usually generate the cluster labels by certain clustering algorithm (e.g., the kmeans) and then perform the matrix approximation guided by such label information. In addition, the noises and outliers in the dataset with large reconstruction errors will easily dominate the objective function by the conventional ℓ_2-norm based squared residue minimization. In this paper, we propose a novel clustering and feature extraction algorithm based on an unified low-rank matrix factorization framework, which suggests that the observed data matrix can be approximated by the production of projection matrix and low dimensional representation, among which the low-dimensional representation can be approximated by the cluster indicator and latent feature matrix simultaneously. Furthermore, we have proposed using the $\ell_{2,1}$-norm and integrating the manifold regularization to further promote the proposed model. A novel Augmented Lagrangian Method (ALM) based procedure is designed to effectively and efficiently seek the optimal solution of the problem. The experimental results in both clustering and feature extraction perspectives demonstrate the superior performance of the proposed method.

Introduction

Low-rank matrix factorization, as a promising technique to find two or more lower dimensional matrices whose product provides a good approximation to the original one and by which to capture the underlying low-dimensional structures of data, plays an important role in many computer vision and pattern recognition applications, e.g., dimension reduction, clustering and classification (Zhang and Zhao 2013; la Torre 2012; Zhang et al. 2015a). As one of the standard approaches for low-rank matrix approximation with a given data matrix and a preindicated rank r of the approximation, the well-known principal component analysis (PCA) is nothing but a truncated singular value decomposition (TSVD) applied on re-centered data (Zhang and Zhao 2013). The low-rank representation (LRR) model (Liu, Lin, and Yu 2010; Liu et al. 2013), which shares the same assumption that the observed data matrix should be approximately of low-rank and seeks the lowest-rank representation of all data jointly, has attracted great deal of attention in recent years and in particular employed for data clustering and segmentation (Liu, Lin, and Yu 2010; Yin et al. 2015). As another famous matrix factorization technique, the nonnegative matrix factorization (NMF) (Lee and Seung 2001) aims to find two nonnegative matrices whose product provides a good approximation to the original one, has also received considerable attention due to its psychological and physiological interpretation of naturally occurring data whose representation may be parts based in the human brain (Cai et al. 2011).

In the literature, the matrix factorization methods can be directly considered as the feature extraction algorithms by letting the factor matrices as the projection matrix and low-dimensional representation, respectively (Guan et al. 2011; 2012; Zhang et al. 2015b). Additional regularizers are accordingly suggested to match the certain data structure or priori knowledge, e.g., the manifold regularization (Zhang and Zhao 2013; Cai et al. 2011), the loss of a classifier (Gupta and Xiao 2011), the model constraint (Chen et al. 2015), and sparsity (Zheng et al. 2012). As an alternate point of view, it have been demonstrated that the NMF is equivalent to the kmeans clustering by interpreting the factor matrices as the cluster indicator and latent feature matrix, respectively (Ding et al. 2005; Ding, Li, and Jordan 2010). Pioneered by this idea, an orthogonal nonnegative matrix tri-factorization algorithm is developed for clustering, which addresses the orthogonality constraint and leads to rigorous clustering interpretation (Ding et al. 2006), and it has been further generalized to a high-order co-clustering framework for simultaneous clustering of multi-type relational data with a fast version to deal with large scale data (Wang et al. 2011). Moreover, an embedded unsupervised feature selection algorithm is proposed by using a novel constraint on the latent feature matrix (Wang, Tang, and Liu 2015).

*Corresponding author.

Despite above achievements on low-rank matrix factorization for data clustering and feature extraction, there are still a few drawbacks for current algorithms: (1) due to the lack of label information in the clustering task, one commonly used criterion is to let the data similarity to be preserved by the predicted labels (He et al. 2011). Although some methods apply sparsity constraints into the matrix factorization to achieve feature dimension reduction (Qian and Zhai 2013), these methods usually generate the cluster labels by certain clustering algorithm (e.g., the kmeans) and then transform unsupervised dimension reduction into the sparsity regularized matrix approximation guided by such cluster labels. (2) In the conventional ℓ_2-norm based squared residue minimization version of matrix factorization, the noises and outliers in the dataset with large reconstruction errors will easily dominate the objective function (Huang et al. 2014; Han et al. 2015). Although the ℓ_1-norm based robust matrix factorization has been proposed to alleviate this issue (Ke and Kanade 2005), it fails to maintain feature rotation invariance. There are also $\ell_{2,1}$-norm based methods for robust NMF (Kong, Ding, and Huang 2011; Huang et al. 2014), but only address the single task of clustering. (3) The standard low-rank matrix factorization ignores the possible nonlinearity inherent in the data. To preserve the local geometrical structures embedded in low-rank matrix factorization, some researchers assume that if two data points are close in the input high-dimensional feature space, then their low-dimensional representations should be close as well (Zhou, Tao, and Wu 2010; Guan et al. 2011; Cai et al. 2011; Zhang and Zhao 2013; Lu et al. 2013). However, it is also essential to assume that if two data points are close in the intrinsic manifold of the data distribution, then their cluster labels should also be close as well (Ng, Jordan, and Weiss 2001).

To relieve above issues in existing low-rank matrix factorization based approaches, in this paper, we propose a novel clustering and feature extraction algorithm based on an unified low-rank matrix factorization framework, i.e., the robust manifold matrix factorization (RMMF). In detail, several highlighted contributions of the proposed approach are summarized as follows.

- We propose a unified low-rank matrix factorization framework that combines clustering and feature extraction in a novel way. In particular, the observed data matrix is approximated by the production of projection matrix and low-dimensional representation, among which the low dimensional representation can be approximated by the cluster indicator and latent feature matrix simultaneously.
- We suggest the $\ell_{2,1}$-norm based matrix factorization in our framework to obtain the robust solution against the noises and outliers. Different from other $\ell_{2,1}$-norm based NMFs for clustering and feature extraction (Huang et al. 2014; Wang, Tang, and Liu 2015), in our method, clustering is performed on the robust low-dimensional representation rather than the input data matrix, which helps our model to better capture the underlying low-dimensional structure and enhance the clustering performance.
- We incorporate the manifold regularization terms on both the low-dimensional feature representation and the cluster labels, to better encode the local geometrical information existing in the data.

The new constraint (i.e., the $\ell_{2,1}$-norm) in our model makes the conventional auxiliary function optimization method no longer applicable for our RMMF problem. Therefore, an Augmented Lagrangian Method (ALM) based procedure is designed to effectively and efficiently seek the optimal solution of the objective function. The rest of the paper is organized as follows: section 2 introduces our RMMF algorithm in detail, section 3 proposes an efficient optimization procedure for RMMF. Then, the experimental results on both clustering and feature extraction perspectives are reported in section 4, followed by the conclusions in section 5.

Robust Manifold Matrix Factorization

Let $X \in \mathrm{R}^{l\times n}$ to be the input data matrix in which n and l are the number of data instances and the original feature dimensionality of each instance, respectively. By a certain linear subspace projection, X can be low-dimensional represented as $Y = P^{\mathrm{T}}X$, in which $P \in \mathrm{R}^{l\times d}$ and $Y \in \mathrm{R}^{d\times n}$ are the projection matrix and the low-dimensional feature representation, respectively. Now we consider the PCA based low-rank matrix factorization under a least-squares framework (la Torre 2012), denote $x_i \in \mathrm{R}^l$ and $y_i \in \mathrm{R}^d$ as instances of X, Y in vector form, respectively, then we have $y_i = P^{\mathrm{T}}x_i$, and the PCA minimizes the following reconstruction error by using the optimal orthogonal basis:

$$\varepsilon = \sum\nolimits_{i=1}^{N} \left\|x_i - P(P^{\mathrm{T}}x_i)\right\|_2^2, \tag{1}$$

in which P is a subset of orthogonal basis of X. Eq. (1) has its matrix formulation as:

$$\varepsilon = \left\|X - P(P^{\mathrm{T}}X)\right\|_F^2. \tag{2}$$

By minimizing the approximation error with a preindicated subspace dimensionality d, the objective of PCA based low-rank matrix approximation can be rewritten as following:

$$\arg\min_{P,Y} \|X - PY\|_F^2, \quad \text{s.t.} \quad P^{\mathrm{T}}P = I. \tag{3}$$

Eq. (3) gives the low-rank matrix approximation of a data matrix based on the feature extraction point of view, in which P is the projection matrix for feature mapping. As an alternate point of view, a data matrix can be clustered into k clusters under an NMF based matrix factorization framework (Ding et al. 2005; Ding, Li, and Jordan 2010). In this paper, we propose to perform such low-rank matrix approximation on the low-dimensional representation Y, but with the relaxed orthogonality constraint on U (Tang and Liu 2012):

$$\begin{aligned} \arg\min_{P,Y,U,V} & \|X - PY\|_F^2 + \left\|Y - VU^{\mathrm{T}}\right\|_F^2, \\ & \text{s.t.} \quad P^{\mathrm{T}}P = I, U^{\mathrm{T}}U = I, U \geq 0, \end{aligned} \tag{4}$$

where $U \in \mathrm{R}^{n\times k}$ is the cluster indicator and $V \in \mathrm{R}^{d\times k}$ is the latent feature matrix (or the cluster centroid). Moreover, in this paper, we propose to further add a $\ell_{2,1}$-norm on V to perform feature selection from Y and enhance the robustness of our model (Nie et al. 2010):

$$\arg\min_{P,Y,U,V} \|X - PY\|_F^2 + \left\|Y - VU^{\mathrm{T}}\right\|_F^2 + \beta \|V\|_{2,1} \quad \text{s.t.} \;\; P^{\mathrm{T}}P = I, U^{\mathrm{T}}U = I, U \geq 0, \tag{5}$$

in which the $\ell_{2,1}$-norm of a matrix A is defined as $\|A\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} A_{ji}^2} = \sum_{i=1}^{n} \|A_i\|$, in order to enforce some columns of V close to 0 and achieve the feature selection from Y.

In the objective function eq. (5), the errors for each data point enters the approximations $\|X - PY\|_F^2$ and $\left\|Y - VU^{\mathrm{T}}\right\|_F^2$ are squared residue errors in the form of ℓ_2-norm. Therefore, the noises and outliers in the dataset with large reconstruction errors will easily dominate the objective function because of the squared errors. Note in our proposed low-rank matrix approximation, both P, Y and U, V are unknown, thus the impact of the outliers may be more complicate than the simpler convex case. To make our model robust to these instances, we replace the loss functions in eq. (5) by the $\ell_{2,1}$-norm as $\|X - PY\|_{2,1}$ and $\left\|Y - VU^{\mathrm{T}}\right\|_{2,1}$, respectively. In this robust matrix factorization, the errors for each data point are $\|x_i - Py_i\|$ and $\left\|y_i - VU_i^{\mathrm{T}}\right\|$, respectively, which is not squared, and thus the large errors due to the noises and outliers in the dataset do not dominate the objective function because they are not squared (Kong, Ding, and Huang 2011). Therefore, we have the following robust low-rank matrix factorization:

$$\arg\min_{P,Y,U,V} \|X - PY\|_{2,1} + \left\|Y - VU^{\mathrm{T}}\right\|_{2,1} + \beta \|V\|_{2,1} \quad \text{s.t.} \;\; P^{\mathrm{T}}P = I, U^{\mathrm{T}}U = I, U \geq 0. \tag{6}$$

Finally, we expect similar data instances from original data matrix X should have similar low-dimensional representation as well as clustering labels, according to the spectral analysis (von Luxburg 2007). Therefore, we incorporate the manifold regularization terms on both the low-dimensional feature representation and the cluster labels, to better encode the local geometrical information existing in the data:

$$\arg\min_{Y,U} \mathrm{tr}(YLY^{\mathrm{T}}) + \mathrm{tr}(U^{\mathrm{T}}LU), \tag{7}$$

where $L = D - W$ is the Laplacian matrix and D is a diagonal matrix with its elements defined as $D_{ii} = \sum_{i=1}^{n} W_{ii}$, and $W \in \mathrm{R}^{n \times n}$ is the relation matrix of X weighted by the RBF kernel (Belkin and Niyogi 2003):

$$W_{ij} = e^{\frac{-\|x_i - x_j\|^2}{\sigma^2}}. \tag{8}$$

Putting eq. (6) and eq. (7) together, the proposed objective function of RMMF is:

$$\begin{aligned} \arg\min_{P,Y,U,V} & \|X - PY\|_{2,1} + \left\|Y - VU^{\mathrm{T}}\right\|_{2,1} \\ & + \alpha(\mathrm{tr}(YLY^{\mathrm{T}}) + \mathrm{tr}(U^{\mathrm{T}}LU)) + \beta \|V\|_{2,1} \\ & \text{s.t.} \;\; P^{\mathrm{T}}P = I, U^{\mathrm{T}}U = I, U \geq 0. \end{aligned} \tag{9}$$

RMMF Optimization

The objective function in above eq. (9) is not convex in four variables but is convex if we update the four variables alteratively. Thus, we use Augmented Lagrangian Method (ALM) to optimize the objective function. By introducing four auxiliary variables $E_1 = X - PY$, $E_2 = Y - VU^{\mathrm{T}}$, $Z_1 = Y$ and $Z_2 = U$. The objective function can be rewritten into the following equivalent problem:

$$\begin{aligned} \arg\min_{P,Y,U,V,E_1,E_2,Z_1,Z_2} & \|E_1\|_{2,1} + \|E_2\|_{2,1} \\ & + \alpha(\mathrm{tr}(Z_1LY^{\mathrm{T}}) + \mathrm{tr}(Z_2^{\mathrm{T}}LU)) + \beta \|V\|_{2,1} \\ & \text{s.t.} \;\; E_1 = X - PY, E_2 = Y - VU^{\mathrm{T}}, Z_1 = Y \\ & \quad Z_2 = U, P^{\mathrm{T}}P = I, U^{\mathrm{T}}U = I, Z_2 \geqslant 0, \end{aligned} \tag{10}$$

which can be solved by the following ALM problem:

$$\begin{aligned} \arg\min_{P,Y,U,V,E_1,E_2,Z_1,Z_2,\lambda_1,\lambda_2,\lambda_3,\lambda_4,\mu} & \|E_1\|_{2,1} + \|E_2\|_{2,1} \\ & + \alpha(\mathrm{tr}(Z_1LY^{\mathrm{T}}) + \mathrm{tr}(Z_2^{\mathrm{T}}LU)) + \beta \|V\|_{2,1} \\ & + < \lambda_1, X - PY - E_1 > + < \lambda_2, Y - VU^{\mathrm{T}} - E_2 > \\ & + < \lambda_3, Z_1 - Y > + < \lambda_4, Z_2 - U > \\ & + \frac{\mu}{2}(\|Z_1 - Y\|_F^2 + \|Z_2 - U\|_F^2 \\ & + \|X - PY - E_1\|_F^2 + \left\|Y - VU^{\mathrm{T}} - E_2\right\|_F^2) \\ & \text{s.t.} \;\; P^{\mathrm{T}}P = I, U^{\mathrm{T}}U = I, Z_2 \geqslant 0, \end{aligned} \tag{11}$$

where λ_1, λ_2, λ_3, and λ_4 are the Lagrangian multipliers and μ is a regularity coefficient to control the penalty for the four violation of equality constraints in eq. (11). Since the objective function above carries eight variables and additional multipliers, we adopt an alternative optimization method to reduce it to a few manageable subproblems with the closed-form solution, each minimizes the objective function with respect to one variable while fixing the other variables. The detailed information is given in Appendix.

Experimental Analysis

In this section, we evaluate the performance of the proposed RMMF method on the benchmark datasets (Tables 1 and 2). We divide this section into two parts to report the experimental results of clustering and feature extraction, respectively (Tables 3 and 4). In each subsection, we firstly introduce the datasets and experimental settings, then compare the proposed RMMF with the state-of-the-art algorithms in detail.

Table 1: Description of datasets for clustering.

Dataset	Classes	Samples	Features
PIE10P	10	210	2420
COIL20	20	1440	1024
USPS	10	930	256
Movement	15	360	90
Glass	6	214	9
Seeds	3	210	7

Table 2: Description of datasets for feature extraction.

Dataset	Classes	Samples	Features
Yale	15	165	1024
ORL	20	1440	1024
MINST	10	930	256
Iohosphere	15	360	34
Sonar	6	214	60
Landsat	3	210	36

Clustering

The clustering experiments are conducted on 6 publicly available benchmark datasets, including a face image dataset (PIE10P), an object image dataset (COIL20), a digit dataset (a subset of USPS), and three non-image datasets from the UCI machine learning repository, i.e., Movement, Glass, and Seeds. The statistics of the datasets used in the clustering experiments are summarized in Table 1.

As indicated in our RMMF algorithm, in this subsection, we consider U as the cluster indicator for clustering. We compare the RMMF with the following representative data clustering algorithms: (1) kmeans in the original input feature space, (2) graph regularized nonnegative matrix factorization (GNMF) (Cai et al. 2011), (3) embedded unsupervised feature selection (EUFS) (Wang, Tang, and Liu 2015), (4) clustering with adaptive neighbors (CAN) (Nie, Wang, and Huang 2014), and (5) sparse manifold clustering and embedding (SMCE) (Elhamifar and Vidal 2011). In all the clustering methods, we set the number of clusters equal to the ground truth class number for all the datasets. To fairly compare different methods, we tune the regularization parameters for all methods by a "grid-search" strategy from the same range of $10^{[-5,-4,\ldots,5]}$. In addition, the EUFS, SMCE, and RMMF are joint dimension reduction and clustering algorithms, which require the subspace dimensionality d as an input, in the experiments, we tune this parameter by using the candidate values which are no more than $l/2$ for various datasets respectively, and report the best performance. The accuracy (ACC) is employed as evaluation metrics to evaluate the performance of clusters (He et al. 2011), the larger scores suggest the better clustering performance. It is also worth noting that the kmeans, GNMF, and EUFS are depend on initialization in optimization, following previous works, we repeat the related experiments ten times and the average results with standard deviation are reported.

Table 3 summarizes the clustering performance for each method on six datasets. We can see that RMMF algorithm outperforms other clustering methods in ACC. In particular, the PIE dataset is a well known dataset that contains a lot of occluded and light-varying images, thus it is often used for robust face recognition (Wright et al. 2009). The significant performance on all these datasets, especially the PIE dataset, meets the major advantages of our method. In detail, comparing to the state-of-the-art data clustering methods, the superiority of the proposed RMMF algorithm lies in the following: (1) the objective function uses the unified low-rank matrix factorization on the input data matrix and the low-dimensional feature representation, which can better capture the underlying low-dimensional structure and enhance the clustering performance, (2) the adopted $\ell_{2,1}$-norm based matrix factorization alleviates the outlier issue that is common among other clustering methods such as the kmeans and GNMF, (3) the manifold regularization terms which incorporate the geometric and manifold information on both low-dimensional feature representation and cluster labels further promote the clustering performance and robustness of the RMMF model. In addition, as pointed in table 3 and many literatures, most of the conventional clustering algorithms (e.g., kmeans, NMF based methods, and EUFS) suffered from the fact of uncertain initialization in optimization, which makes the clustering results difficult to be exactly reproduced. On the contrary, the performance of our proposed RMMF clustering is efficient and stable by our suggested optimization procedure.

Feature Extraction

We would like to show that our RMMF model also applies to feature extraction task and benefits for subsequent classification. In this subsection, there are six datasets used for experimental evaluation, including two face image ones: Yale and ORL, a digit dataset (MINST), and other three non-image datasets from the UCI machine learning repository, i.e., Iohosphere, Sonar, and Landsat. Again, the information about the number of classes, samples, and features in each dataset is given in Table 2.

As indicated in our RMMF algorithm, in this subsection, we consider P as the projection matrix for feature extraction. We provide experimental results on the classification task of datasets above to test the performance of our proposed RMMF algorithm, and compare it to the state-of-the-art feature dimension reduction methods including the PCA, kernel PCA (KPCA), locality preserving projections (LPP) (He and Niyogi 2004), and neighborhood preserving embedding (NPE) (He et al. 2005). For each algorithm, after the feature embedding is obtained, a lazy classifier, i.e., the k-nearest-neighbor with k=1 is used for classification. Also, the kNN classification using the original high-dimensional feature representation is performed as a baseline for all the feature extraction methods.

In the experiment, the parameter setting of RMMF is the same as mentioned in the above subsection, while for LPP and NPE, we search the parameter k in the range of [2,4,...,20] and the parameter t in the range of $10^{[-3,-2,\ldots,3]}$ for LPP. The dimensionality of feature embedding d is critical in feature extraction algorithms, in our experiments, we tune it by using different candidate sets according to the original feature size of each dataset and show the best

Table 3: Clustering results of different methods by the measurement of ACC in percentage.

Dataset	kmeans	GNMF	EUFS	CAN	SMCE	RMMF
PIE10P	29.90±2.43	41.19±2.36	65.71±3.56	52.38	69.05	**73.81**
COIL20	54.94±4.10	49.44±3.62	60.23±4.04	84.58	71.60	**86.94**
USPS	62.62±4.62	67.31±3.63	68.06±5.44	70.00	64.73	**76.56**
Movement	43.64±2.88	40.28±2.71	49.44±5.11	51.67	52.66	**54.44**
Glass	50.79±3.98	54.67±2.91	56.54±3.11	51.40	56.07	**61.68**
Seeds	72.14±6.48	84.29±4.10	86.67±0.49	88.10	87.62	**90.48**

Table 4: Feature extraction and classification results of different methods by the measurement of OA in percentage.

Dataset	kNN	PCA	KPCA	LPP	NPE	RMMF
Yale	63.33 (1024)	66.67 (100)	68.89 (30)	70.00 (20)	67.78 (30)	**74.44** (50)
ORL	76.43 (1024)	76.79 (100)	77.14 (100)	79.64 (30)	80.36 (30)	**82.85** (150)
MINST	68.00 (256)	73.00 (20)	**75.56** (30)	75.22 (20)	71.56 (10)	73.56 (30)
Iohosphere	73.41 (34)	75.23 (8)	76.13 (4)	73.41 (8)	74.92 (16)	**81.27** (10)
Sonar	61.17 (60)	62.23 (30)	64.36 (10)	65.96 (20)	69.15 (10)	**73.94** (22)
Landsat	69.64 (36)	73.09 (4)	76.29 (18)	71.65 (6)	72.01 (6)	**73.56** (12)

performance along with the best number of d, the detailed candidate sets are as following: [10:10:150], [10:10:150], [10:10:100], [2:2:16], [2:2:30] and [2:2:18], for six datasets, respectively. Finally, for the kNN classification, we select the first [5, 3, 10, 10, 10, 10] samples for each class as training set for six datasets, respectively, and the other samples are left for testing.

The classification overall accuracies (OA) of all the feature embedding algorithms are reported in Table 4. In the table, the number in parentheses is the number of features when the best performance is achieved. From this table we learn that the classification accuracies using the extracted features are always improved compare to the original kNN. We also find that most of the time, the proposed RMMF algorithm outperforms other methods in more than 3 percentages, which demonstrates the effectiveness of our RMMF algorithm for feature extraction.

Conclusion

In this paper, we propose a robust manifold matrix factorization (RMMF) for joint clustering and feature extraction. The RMMF is an unified low-rank matrix factorization framework which combines clustering and feature extraction in a novel way, furthermore, the $\ell_{2,1}$-norm is applied to the matrix factorization to obtain the robust solution against the noises and outliers, and the manifold regularization term is introduced to better incorporate the geometric and manifold information on both low-dimensional feature representation and cluster labels. The proposed model can perform both data clustering via U and feature extraction via P, experimental results on numerous of datasets (including face, object, and digit image datasets and other non-image datasets) demonstrate the superior performance of the proposed method in both clustering and feature extraction perspectives. For future work, the proposed method can be further extended to more challenging tasks such as multi-view and cross-view data clustering and feature extraction.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 61401317, 61471274 and 91338202, the Australian Research Council Projects DP-140102164, FT-130101457, and LE-140100061, and the Hong Kong Scholars Program and Hong Kong Government General Research Fund Scheme under grant 152202/14E.

Appendix

To optimize eq. (11), the following steps are repeated until convergence.

Update E_1

To update E_1, we fix other variables except E_1 and remove terms that are irrelevant to E_1. Then eq. (11) becomes:

$$\arg\min_{E_1} \frac{1}{\mu} \|E_1\|_{2,1} + \frac{1}{2} \left\| E_1 - (X - PY + \frac{1}{\mu}\lambda_1) \right\|_F^2 . \tag{12}$$

This equation has a closed form solution (Liu, Ji, and Ye 2009). Let $B = X - PY + \frac{1}{\mu}\lambda_1$, then E_1 can be updated as:

$$E_{1i} = \begin{cases} (1 - \frac{1}{\mu\|B_i\|})B_i, & \text{if } \|B_i\| \geq \frac{1}{\mu} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Update E_2

To update E_2, we fix other variables except E_2 and remove terms that are irrelevant to E_2. Then eq. (11) becomes:

$$\arg\min_{E_2} \frac{1}{\mu} \|E_2\|_{2,1} + \frac{1}{2} \left\| E_2 - (Y - VU^{\mathrm{T}} + \frac{1}{\mu}\lambda_2) \right\|_F^2 . \tag{14}$$

Similar to above step of update E_1, we let $C = Y - VU^{\mathrm{T}} + \frac{1}{\mu}\lambda_2$, then E_2 can be updated as:

$$E_{2i} = \begin{cases} (1 - \frac{1}{\mu\|C_i\|})C_i, & \text{if } \|C_i\| \geq \frac{1}{\mu} \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Update V

To update V, we fix other variables except V, then the objective function in eq. (11) reduces to:

$$\arg\min_{V} \beta \|V\|_{2,1} + \frac{\mu}{2}\left\|Y - VU^{\mathrm{T}} - E_2 + \frac{1}{\mu}\lambda_2\right\|_F^2, \tag{16}$$

By considering the constraint of $U^{\mathrm{T}}U = I$, we can rewrite it as:

$$\arg\min_{V} \frac{\beta}{\mu} \|V\|_{2,1} + \frac{1}{2}\left\|V - (Y - E_2 + \frac{1}{\mu}\lambda_2)U\right\|_F^2, \tag{17}$$

Similar to above step of update E_1, if we denote $M = (Y - E_2 + \frac{1}{\mu}\lambda_2)U$, then we have:

$$V_i = \begin{cases} (1 - \frac{\beta}{\mu\|M_i\|})M_i, & \text{if } \|M_i\| \geq \frac{\beta}{\mu} \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Update Y

Optimizing eq. (11) with respect to Y yields the equation:

$$\begin{aligned} &\arg\min_{Y} < \lambda_1, X - PY - E_1 > + < \lambda_2, Y - VU^{\mathrm{T}} - E_2 > \\ &+ \frac{\mu}{2}(\|Z_1 - Y\|_F^2 + \|X - PY - E_1\|_F^2 \\ &+ \left\|Y - VU^{\mathrm{T}} - E_2\right\|_F^2) + \alpha \mathrm{tr}(Z_1 L Y^{\mathrm{T}}) \end{aligned} \tag{19}$$

By considering the constraint of $P^{\mathrm{T}}P = I$, we can reformulate the above function as:

$$\begin{aligned} &\arg\min_{Y} < \lambda_1, X - PY - E_1 > + < \lambda_2, Y - VU^{\mathrm{T}} - E_2 > \\ &+ \frac{\mu}{2}(\|Z_1 - Y\|_F^2 + \left\|Y - P^{\mathrm{T}}(X - E_1)\right\|_F^2 \\ &+ \left\|Y - VU^{\mathrm{T}} - E_2\right\|_F^2) + \alpha \mathrm{tr}(Z_1 L Y^{\mathrm{T}}) \end{aligned} \tag{20}$$

By letting the Lagrangian function of eq. (20) to 0, we have:

$$\begin{aligned} Y = &\frac{1}{3\mu}((P^{\mathrm{T}}\lambda_1 - \lambda_2 + \mu(Z_1 + P^{\mathrm{T}}(X - E_1) \\ &+ (VU^{\mathrm{T}} - E_2)) - \alpha Z_1 L)). \end{aligned} \tag{21}$$

Update Z_1

The proposed objective function with respect to Z_1 yields the equation:

$$\arg\min_{Z_1} \alpha \mathrm{tr}(Z_1 L Y^{\mathrm{T}}) + < \lambda_3, Z_1 - Y > + \frac{\mu}{2}(\|Z_1 - Y\|_F^2), \tag{22}$$

by letting the Lagrangian function of eq. (22) to 0, we have:

$$Z_1 = \frac{\mu Y - \lambda_3 - \alpha Y L}{\mu} \tag{23}$$

Update Z_2

Optimizing Equation eq. (11) with respect to Z_2 is reduced to the following equation:

$$\arg\min_{Z_2 \geq 0} \alpha \mathrm{tr}(Z_2^{\mathrm{T}} L U) + < \lambda_4, Z_2 - U > + \frac{\mu}{2}\|Z_2 - U\|_F^2, \tag{24}$$

which can be further reduced as following:

$$\arg\min_{Z_2 \geq 0} \|Z_2 - K\|_F^2, \tag{25}$$

where $K = (U - \frac{1}{\mu}\lambda_4 - \frac{\alpha}{\mu}LU)$. Eq. (25) can be further decomposed to element-wise optimization problem as:

$$\arg\min_{Z_{2ij} \geq 0} \left\|Z_{2ij} - K_{ij}\right\|_F^2. \tag{26}$$

Therefore, the optimal solution of Z_2 should be:

$$Z_{2ij} = \max(K_{ij}, 0). \tag{27}$$

Update P

Optimizing eq. (11) with respect to P yields the equation:

$$\arg\min_{P^{\mathrm{T}}P = I} < \lambda_1, X - PY - E_1 > + \frac{\mu}{2}(\|X - PY - E_1\|_F^2), \tag{28}$$

Eq. (28) can be further rewritten as:

$$\arg\min_{P^{\mathrm{T}}P = I} \frac{\mu}{2}\left\|X - PY - E_1 + \frac{1}{\mu}\lambda_1\right\|_F^2. \tag{29}$$

If we define $\Theta = (X - E_1 + \frac{1}{\mu}\lambda_1)Y^{\mathrm{T}}$, then eq. (29) equals to:

$$\arg\min_{P^{\mathrm{T}}P = I} \|P - \Theta\|_F^2. \tag{30}$$

which can be solved as $P = N_p Q_p^{\mathrm{T}}$ in which N_p and Q_p are the left and right singular vectors of the singular value decomposition of Θ (Huang et al. 2014).

Update U

The proposed objective function with respect to U yields the equation:

$$\begin{aligned} &\arg\min_{U^{\mathrm{T}}U = I} \alpha \mathrm{tr}(Z_2^{\mathrm{T}} L U) + < \lambda_4, Z_2 - U > \\ &+ < \lambda_3, Y - VU^{\mathrm{T}} - E_2 > \\ &+ \frac{\mu}{2}(\|Z_2 - U\|_F^2 + \left\|Y - VU^{\mathrm{T}} - E_2\right\|_F^2). \end{aligned} \tag{31}$$

Then, by removing the irrelevant terms and defining $\Psi = \frac{1}{\mu}\lambda_4 + Z_2^{\mathrm{T}} - \frac{\beta}{\mu}LZ_2 + (Y - E_2)^{\mathrm{T}}(V - \frac{1}{\mu}\lambda_3)$, eq. (31) arrives at:

$$\min_{U^{\mathrm{T}}U = I} \frac{\mu}{2}\|U\|_F^2 - \mu < \Psi, U > \tag{32}$$

and it can be further simplified as:

$$\min_{U^{\mathrm{T}}U = I} \|U - \Psi\|_F^2. \tag{33}$$

Similar to the solution of eq. (30), we have $U = N_u Q_u^{\mathrm{T}}$ where N_u and Q_u are the left and right singular vectors of the singular value decomposition of Ψ.

Update ALM Parameters

Finally we need to update the ALM parameters, i.e., λ_1, λ_2, λ_3, λ_4 and μ. According to (Boyd and Vandenberghe 2004), they should be updated as following:

$$\lambda_1 = \lambda_1 + \mu(X - PY^{\mathrm{T}} - E_1) \tag{34}$$

$$\lambda_2 = \lambda_2 + \mu(Y - UV^{\mathrm{T}} - E_2) \tag{35}$$

$$\lambda_3 = \lambda_3 + \mu(Z_1 - Y) \tag{36}$$

$$\lambda_4 = \lambda_4 + \mu(Z_2 - U) \tag{37}$$

$$\mu = \rho\mu \tag{38}$$

References

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15(6):1373–1396.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI* 33(8):1548–1560.

Chen, P.; Wang, N.; Zhang, N. L.; and Yeung, D.-Y. 2015. Bayesian adaptive matrix factorization with automatic model selection. In *Proc. CVPR*, 1284–1292.

Ding, C.; He, X.; Simon, H. D.; and Jin, R. 2005. On the equivalence of nonnegative matrix factorization and k-means-spectral clustering. In *Proc. SIAM Conf Data Mining*, 606–610.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proc. SIGKDD*, 126–135.

Ding, C.; Li, T.; and Jordan, M. I. 2010. Convex and semi-nonnegative matrix factorizations. *IEEE TPAMI* 32(1):45–55.

Elhamifar, E., and Vidal, R. 2011. Sparse manifold clustering and embedding. In *Proc. NIPS*, 977–986.

Guan, N.; Tao, D.; Luo, Z.; and Yuan, B. 2011. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE TIP* 20(7):2030–2048.

Guan, N.; Tao, D.; Luo, Z.; and Yuan, B. 2012. Nenmf: An optimal gradient method for non-negative matrix factorization. *IEEE TSP* 60(6):2882–2898.

Gupta, M. D., and Xiao, J. 2011. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In *Proc. CVPR*, 2841–2848.

Han, J.; Zhang, D.; Hu, X.; Guo, L.; Ren, J.; and Wu, F. 2015. Background prior-based salient object detection via deep reconstruction residual. *IEEE TCSVT* 25(8):1309–1321.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *Proc. NIPS*, 153–160.

He, X.; Cai, D.; Yan, S.; and Zhang, H.-J. 2005. Neighborhood preserving embedding. In *Proc. ICCV*, 1208–1213.

He, X.; Cai, D.; Shao, Y.; Bao, H.; and Han, J. 2011. Laplacian regularized gaussian mixture model for data clustering. *IEEE TKDE* 23(9):1406–1418.

Huang, J.; Nie, F.; Huang, H.; and Ding, C. 2014. Robust manifold nonnegative matrix factorization. *ACM TKDD* 8(3):1–21.

Ke, Q., and Kanade, T. 2005. Robust ℓ_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proc. CVPR*, 774–783.

Kong, D.; Ding, C.; and Huang, H. 2011. Robust nonnegative matrix factorization using $\ell_{2,1}$-norm. In *Proc. CIKM*, 673–682.

la Torre, F. D. 2012. A least-squares framework for component analysis. *IEEE TPAMI* 34(6):1041–1055.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Proc. NIPS*, 556–562.

Liu, G.; Lin, Z.; Yan, S.; Sun, J.; Yu, Y.; and Ma, Y. 2013. Robust recovery of subspace structures by low-rank representation. *IEEE TPAMI* 35(1):171–184.

Liu, J.; Ji, S.; and Ye, J. 2009. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In *Proc. UAI*, 339–348.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *Proc. ICML*, 663–670.

Lu, X.; Wu, H.; Yuan, Y.; Yan, P.; and Li, X. 2013. Manifold regularized sparse nmf for hyperspectral unmixing. *IEEE TGRS* 51(5):2815–2826.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Proc. NIPS*, 849–856.

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

Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *Proc. SIGKDD*, 977–986.

Qian, M., and Zhai, C. 2013. Robust unsupervised feature selection. In *Proc. IJCAI*, 1621–1627.

Tang, J., and Liu, H. 2012. Unsupervised feature selection for linked social media data. In *Proc. SIGKDD*, 904–912.

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Wang, H.; Nie, F.; Huang, H.; and Ding, C. 2011. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation. In *Proc. ICDM*, 774–783.

Wang, S.; Tang, J.; and Liu, H. 2015. Embedded unsupervised feature selection. In *Proc. AAAI*, 470–476.

Wright, J.; Yang, A. Y.; Ganesh, A.; Sastry, S. S.; and Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE TPAMI* 31(2):210–227.

Yin, M.; Gao, J.; Lin, Z.; Shi, Q.; and Guo, Y. 2015. Dual graph regularized latent low-rank representation for subspace clustering. *IEEE TIP* 24(12):4918–4933.

Zhang, Z., and Zhao, K. 2013. Low-rank matrix approximation with manifold regularization. *IEEE TPAMI* 35(7):1717–1729.

Zhang, L.; Zhang, L.; Tao, D.; Huang, X.; and Du, B. 2015a. Compression of hyperspectral remote sensing images by tensor approach. *Neurocomputing* 147:358–363.

Zhang, L.; Zhang, Q.; Zhang, L.; Tao, D.; Huang, X.; and Du, B. 2015b. Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding. *Pattern Recognit.* 48(10):3102–3112.

Zheng, Y.; Liu, G.; Sugimoto, S.; Yan, S.; and Okutomi, M. 2012. Practical low-rank matrix approximation under robust l1-norm. In *Proc. CVPR*, 1410–1417.

Zhou, T.; Tao, D.; and Wu, X. 2010. Manifold elastic net: A unified framework for sparse dimension reduction. *Data Min. Knowl. Disc.* 22(3):340–371.