2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Linear Subspace Learning via Sparse Dimension Reduction

Ming Yin
School of Automation
Guangdong University of Technology
Guangzhou, China.
Email: yiming@gdut.edu.cn

Yi Guo
CSIRO Mathematics
Informatics and Statistics
North Ryde, NSW 1670, Australia.
Email:yi.guo@csiro.au

Junbin Gao
School of Computing and Mathematics
Charles Sturt University
Bathurst, NSW 2795, Australia.
Email: jbgao@csu.edu.au

***Abstract*—Linear Subspace Learning (LSL) has been widely used in many areas of information processing, such as dimensionality reduction, data mining, pattern recognition and computer vision. Recent years have witnessed several excellent extensions of PCA in LSL. One is the recent L1-norm maximization principal component analysis (L1Max-PCA), which aims at learning linear subspace efficiently. L1Max-PCA simply simulates PCA by replacing the covariance with the so-called L1-norm dispersion in the mapped feature space. However, it is difficult to give an intuitive interpretation. In this paper, a novel subspace learning approach based on sparse dimension reduction is proposed, which enforces the sparsity of the mapped data to better recover cluster structures. The optimization problem is solved efficiently via Alternating Direction Method (ADM). Experimental results show that the proposed method is effective in subspace learning.**

***Index terms*—** subspace learning; L1-norm; principal component analysis (PCA); Alternating Direction Method

I. Introduction

Automated learning of low-dimensional linear or multilinear models from training data has become a standard paradigm in machine learning. This will greatly benefit describing class relationships among observed objects with lower misclassification rates [14]. However, as observed data are usually embedded in a high-dimensional space, one has to confront the increasing computational complexity of subsequent tasks in high dimension, which is commonly referred to as the curse of dimensionality [6]. Moreover, the classification performance could be very poor in the cases where only limited number of data are available.

In the past decades, linear subspace learning (LSL) has been considered as a powerful tool for dimensionality reduction, which has been widely used in many applications [4], such as image segmentation [24], motion segmentation [22], face clustering [13] and image processing [12]. By projecting data onto the learned subspace, LSL can effectively reduce the dimensionality of input data to simplify subsequent processing tasks without degrading too much performance. Recently, many LSL methods have been proposed such as null-space linear discriminant analysis(NLDA) [5], Locality Preserving Projections (LPP) [2], Marginal Fisher Analysis (MFA) [23], and many more.

LSL is regarded as a pre-processing step in data analysis, aiming to reduce the computational complexity for subsequent processing while maintaining a minimal loss of information. This is usually achieved by optimizing some criterion function, e.g. least squares of error. For example, Zhang *et al.* [25] proposed a novel LSL approach by using sparse coding and feature grouping. The classical PCA tries to find a set of projections that maximize the covariance of the projected data [16]. This is equivalent to minimizing reconstruction error measured by L2 norm, i.e. least squares. So we call it L2 norm PCA. The L2-norm PCA has been widely and successfully applied in data analysis [17], However, it is well known that the L2-norm PCA is sensitive to outliers due to the least squares of error. The L1-norm PCA [15] was proposed to alleviate this disadvantage by applying maximum likelihood estimation to input data. By combining the advantages of L2-norm PCA and L1-norm PCA, Ding *et al.* proposed R1-PCA [7] to suppress the effect of outliers. Although R1-PCA is robust to outliers, the algorithm relies on the knowledge of the dimension of the subspace to be learned. A more sophisticated L1-PCA was proposed by using a probabilistic framework and the model realization is implemented by variational Bayesian [10]. Motivated by the fact that data often exhibit some sparsity, Zou *et al.* [26] proposed an elegant sparse PCA algorithm (SPCA) using the elastic net for L1-penalized regression on regular principle components.

In order to achieve robustness and rotational invariance, Kwak [16] proposed a novel PCA recently under the criterion of maximizing L1-norm of the projection of data in feature space called L1Max-PCA. The subspace learning problem formulated by L1-norm maximization is handled by greedy search. However, a greedy search algorithm is often trapped in a local optimum. To achieve a possible global solution, a non-greedy L1-norm maximization was proposed for robust principal component analysis [20].

A drawback of L1Max-PCA is that it lacks an intuitive interpretation of the L1 dispersion. In this paper, instead of maximizing L1-norm in feature space, we minimize the L1-norm of the dispersion. The motivation is that the minimization of a L1-norm usually produces sparse solutions leading to robustness to outliers. Moreover, the minimization of the L1-norm in the learned subspace gives an intuitive interpretation, i.e. encouraging sparsity of projected data. We explain the proposed method in the rest of this paper. In Section II, we briefly review the PCA-based subspace learning methods and explain the motivation of our approach. The L1-norm minimization PCA for subspace learning is detailed in Section III. In Section IV, we propose an optimization algorithm

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

based on ADM. To evaluate the proposed method, we conduct several experiments on real world database in Section V. Finally, some conclusions are summarized in Section VI.

II. Problem Definition

In this section, we give a brief review of the conventional PCA from LSL point of view. Denote a given data matrix by $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$ where n and d are the size and the dimension of samples respectively. We assume that $\mathbf{X}$ has already been centralized. A projection matrix is denoted by $\mathbf{P} \in \mathbb{R}^{d\times m}$, and $\mathbf{p}^i$ is the ith row of $\mathbf{P}$ and $\mathbf{p}_j$ the jth column. The Frobenius norm of $\mathbf{P}$ is denoted as $\|\mathbf{P}\|_F = \text{tr}\left(\mathbf{P}^T\mathbf{P}\right)^{1/2}$ where tr$(\cdot)$ is the trace operator of a matrix.

The traditional PCA, i.e. L2 norm PCA, is to seek an $m(< d)$ dimensional linear subspace by minimizing the reconstruction error measured by L2 norm as follows,

$$\min_{\mathbf{P},\mathbf{V}} \|\mathbf{X} - \mathbf{P}\mathbf{V}\|_2, \text{ s.t.} \mathbf{P}^\mathbf{T}\mathbf{P} = \mathbf{I}, \tag{1}$$

where $\mathbf{V} = \mathbf{P}^\mathbf{T}\mathbf{X}$ is the projection of $\mathbf{X}$ on $\mathbf{P}$ and $\mathbf{I}$ is the identity matrix with compatible dimensions. This problem can be efficiently solved by eigen decomposition. However, it is well known that L2-norm of error is sensitive to outliers as the underlying error distribution in (1) is Gaussian. It has been proved that the L1-norm of error is more robust to outliers, so it is easy to apply L1-norm to the error in (1) as

$$\min_{\mathbf{P},\mathbf{V}} \|\mathbf{X} - \mathbf{P}\mathbf{V}\|_1, \text{ s.t. } \mathbf{P}^\mathbf{T}\mathbf{P} = \mathbf{I}. \tag{2}$$

The L1-norm optimization in (2) improves the robustness to outliers. However, it is variant to rotations and the exact solution is also difficult to obtain. Gao [10] adopted a Bayesian framework to solve it partially.

As a compromise, Ding et al [7] proposed the R1-norm PCA to learn a subspace by solving an R1-norm minimization problem,

$$\min_{\mathbf{P},\mathbf{V}} \|\mathbf{X} - \mathbf{P}\mathbf{V}\|_{R1} \triangleq \sum_{i=1}^{n}\left(\sum_{j=1}^{d}\left(x_{ji} - \sum_{k=1}^{m} p_{jk}v_{ki}\right)^2\right)^{1/2} \tag{3}$$

In fact, R1-norm (also called L2/L1-norm) PCA is a combination of L2-norm PCA and L1-norm PCA. R1-norm PCA is solved by performing a subspace iteration algorithm in the original space, which is computationally expensive.

Interesting enough, L2 norm PCA can also be regarded as finding a projection matrix to maximize the Frobenius norm of the covariance of projected data as the following

$$\max_{\mathbf{P}} \left\|\mathbf{P}^T\mathbf{X}\right\|_F, \quad \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}. \tag{4}$$

This interpretation directly leads to the L1Max-PCA [16] as follows,

$$\max_{\mathbf{P}} \left\|\mathbf{P}^\mathbf{T}\mathbf{X}\right\|_1, \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}. \tag{5}$$

$$\tag{6}$$

This straightforward extension of the L2 norm PCA in (4) is invariant to rotation and robust to outliers. However, an intuitive interpretation of L1Max-PCA is absent. A greedy search [16] was applied to solve (6), which sequentially optimizes projection direction one by one. However, it is well known that greedy search is often trapped into a local optimum. To resolve this issue, Nie [20] proposed a non-greedy scheme by optimizing all the projection directions simultaneously. For more details on greedy algorithms for L1Max-PCA, please refer to [9].

III. L1-norm Minimization PCA

As pointed out in the previous section, one drawback of L1Max-PCA is its lack of direct interpretation of maximizing the L1 dispersion. We henceforth adopt a minimization strategy for the subspace learning. The reason is that minimizing L1-norm of the projected data encourages sparsity in the target space. Actually, it has been found that natural signals, in various computer vision and pattern recognition applications, can be generally represented by a small number of basis functions chosen out of an over-complete code set [21]. This observation gives rise to the work in compressive sensing. The research on sparse methods in the last decade suggests that under a minimization strategy we are seeking for a sparse solution to the projection [19], so that the projected data lie on either one low dimensional subspace or the union of several disjoint subspaces. If data can be well represented in terms of a few coordinates, it will be much easier to analyze and interpret them in subsequent processing. Motivated by this understanding, we propose a novel subspace learning method which minimizes L1-norm of the projected data.

Different from the way that imposes a sparse penalty to the loading matrix [26], we minimize the L1-norm of projected data to enforce the sparsity in the projected space as follows,

$$\min_{\mathbf{P}} \left\|\mathbf{P}^\mathbf{T}\mathbf{X}\right\|_1, \text{s.t. } \mathbf{P}^T\mathbf{P} = \mathbf{I}. \tag{7}$$

Our model aims at finding a projection matrix by which the data can be directly compressed into low dimensional and sparse representations so that the preferred information underlain in the high dimensional and dense features can be preserved subsequently. We hereafter call the proposed approach L1Min PCA.

The orthonormality requirement $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ in (7) constrains the solution over the Grassmann manifold resulting in a difficult problem. Although the orthogonal projection is desired in practice, it does complicate the optimization problem drastically. Introducing some slackness to this stringent condition can greatly simplify the problem. Here we propose the relaxed L1Min PCA as follows

$$\min_{\mathbf{P}} \left\|\mathbf{P}^T\mathbf{X}\right\|_1 + \lambda\left\|\mathbf{P}^T\mathbf{P} - \mathbf{I}\right\|_F^2 \tag{8}$$

In (8), the first term defines the L1-norm of the projected data, which encourages sparsity. The second term is to relax the orthonormality requirement on the projection. λ is a regularization parameter to balance the sparsity and orthonormality.

Algorithm 1: Minimization L1 PCA based on ADM

Input: $X, \lambda > 0, \rho > 1$

Initialization: $\mathbf{G}_1 = 0$; $\mathbf{Q}_1 = X$; $\mu_1 = 0.1$; $\rho = 1.1$; $\lambda = 0.01$; $k = 1$.

Output: an optimal solution of the projection, i.e. $\mathbf{P}^*$.

While not converged **do**

1. solve $\min_{\mathbf{P}} -\langle \mathbf{G}_k, \mathbf{P}^T\mathbf{X}\rangle + \frac{\mu_k}{2}\|\mathbf{Q}_k - \mathbf{P}^T\mathbf{X}\|_F^2 + \lambda\left\|\mathbf{P}^T\mathbf{P} - \mathbf{I}_m\right\|_F^2$ by L-BFGS solver [18],
2. update $\mathbf{Q_{k+1}}$: $\mathbf{Q_{k+1}} = \mathcal{S}_{1/\mu_k}(\mathbf{P}_{k+1}^T\mathbf{X} - \frac{\mathbf{G}_k}{\mu_k})$,
3. update $\mathbf{G}_{k+1}$: $\mathbf{G}_{k+1} = \mathbf{G}_k + \mu_k(\mathbf{Q}_k - \mathbf{P}_k^T\mathbf{X})$,
4. update μ_{k+1}: $\mu_{k+1} = \rho \cdot \mu_k$.
5. $k = k + 1$.

End while

IV. Optimization based on Alternating Direction Method

Although the solution to problem (8) can be solved by gradient descent with projection [18], in this paper, we apply the so-called Alternating Direction Method (ADM) [3] for its high efficiency. ADM is a practical improvement of the classical Augmented Lagrangian method for solving convex programming problems with convex constraints. In recent years, it has been widely used in many applications. Furthermore, because L1-norm is not differentiable, using a smooth approximation to L1-norm is gradient descent brings another layer of complexity. However, by introducing an auxiliary variable $\mathbf{Q}$ such that $\mathbf{Q} = \mathbf{P^T X}$ in ADM, we avoid minimizing L1-norm of $\mathbf{P^T X}$ directly. Thus, (8) is reformulated as

$$\min_{\mathbf{P},\mathbf{Q}} \|\mathbf{Q}\|_1 + \lambda\left\|\mathbf{P}^T\mathbf{P} - \mathbf{I}_m\right\|_F^2, \qquad \text{s.t. } \mathbf{Q} = \mathbf{P}^T\mathbf{X}.$$

The augmented Lagrangian function of the above problem is given by

$$\begin{aligned}\mathcal{L}(\mathbf{P}, \mathbf{Q}, \mathbf{G}, \mu) =& \|\mathbf{Q}\|_1 + \langle \mathbf{G}, \mathbf{Q} - \mathbf{P}^T\mathbf{X}\rangle \\ &+ \frac{\mu}{2}\|\mathbf{Q} - \mathbf{P}^T\mathbf{X}\|_F^2 + \lambda\left\|\mathbf{P}^T\mathbf{P} - \mathbf{I}_m\right\|_F^2,\end{aligned}$$

where $\mathbf{G}$ is the matrix of Lagrange multipliers and $\mu > 0$ is a penalty parameter. Then the alternating direction optimization for L1Min-PCA goes as follows:

$$\mathbf{P}_{k+1} = \operatorname*{argmin}_{\mathbf{P}} \mathcal{L}(\mathbf{P}, \mathbf{Q}_k, \mathbf{G}_k, \mu_k) \tag{9}$$

$$\mathbf{Q}_{k+1} = \operatorname*{argmin}_{\mathbf{Q}} \mathcal{L}(\mathbf{P}_{k+1}, \mathbf{Q}, \mathbf{G}_k, \mu_k) \tag{10}$$

$$\mathbf{G}_{k+1} = \mathbf{G}_k + \mu_k(\mathbf{Q}_k - \mathbf{P}_k^T\mathbf{X}) \tag{11}$$

$$\mu_{k+1} = \rho\mu_k \tag{12}$$

where $\rho > 1$ is a constant. Subproblem (9) is

$$\min_{\mathbf{P}} -\langle \mathbf{G}_k, \mathbf{P}^T\mathbf{X}\rangle + \frac{\mu_k}{2}\|\mathbf{Q}_k - \mathbf{P}^T\mathbf{X}\|_F^2 + \lambda\left\|\mathbf{P}^T\mathbf{P} - \mathbf{I}_m\right\|_F^2,$$

which can be solved by a unconstrained gradient based solver such as Limited-memory BFGS or CG [18]. Subproblem (10) is

$$\min_{\mathbf{Q}} \|\mathbf{Q}\|_1 + \langle \mathbf{G}_k, \mathbf{Q}\rangle + \frac{\mu_k}{2}\|\mathbf{Q} - \mathbf{P}_k^T\mathbf{X}\|_F^2,$$

which is a proximity problem for L1-norm and has a closed-form solution [1] given by the so-called shrinkage (or soft-thresholding) operator defined as follows,

$$\mathcal{S}_\tau(x) = \operatorname{sgn}(x) \times \max(|x| - \tau, 0)$$

where $\tau > 0$.

The complete algorithm is summarized in Algorithm 1.

(a)

(b)

Fig. 1. Samples for performance testing. (a) *ORL* database, (b) *COIL100* database.

V. Experimental Results

In this section, we evaluate the performance of the proposed L1Min-PCA for unsupervised learning and supervised learning tasks by comparing against L2-norm PCA (L2PCA), L1-norm maximization PCA with greedy algorithm [16](L1PCA_G) and its non-greedy version [20] (L1PCA_NG).

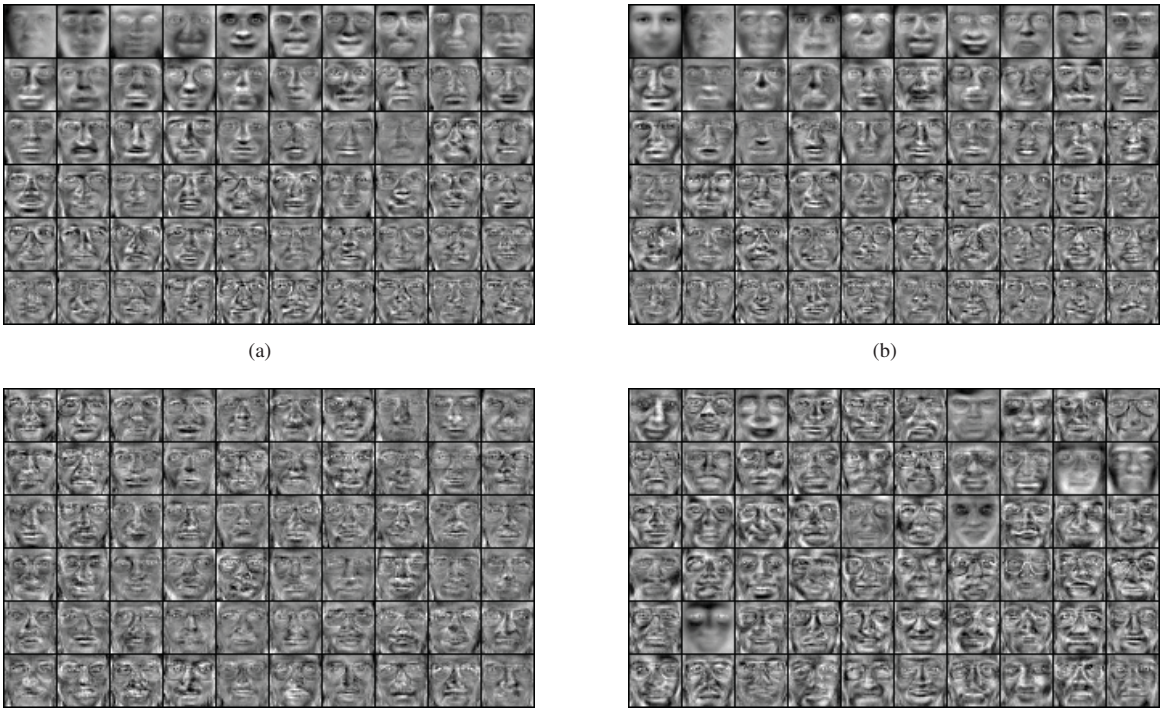
Four public test databases were used as test data in our experiments. The first one is CMU PIE face database[1], which contains 41,368 face images of 68 human subjects. These face images were acquired from 13 synchronized cameras and 21 flashes, with varying poses, illustration and expression. In our experiment, the frontal poses (named as C27) with different illustration and expressions were selected, which contains 3329 face images in total. The second database is the Extended Yale-B[2], consisting of 16128 images of 38 human subjects

[1]http://www.ri.cmu.edu/projects/project_418.html

[2]http://vision.ucsd.edu/ leekc/ExtYaleDatabase/ExtYaleB.html

TABLE I
CLASSIFICATION ERROR RATES (%) BY DIFFERENT METHODS

Method	PIE						YaleB					
	33% Test		50%Test		66% Test		33% Test		50%Test		66% Test	
	Dim.	error	Dim.	error	Dim.	error	Dim.	error	Dim.	error	Dim.	error
Baseline	1024	11.17	1024	6.73	1024	3.06	1024	28.64	1024	19.47	1024	16.27
L2PCA	650	11.17	650	6.73	650	3.06	850	28.63	850	19.55	900	16.27
L1PCA_G	325	11.70	400	6.78	340	3.15	185	31.6	185	23.2	185	19.9
L1PCA_NG	400	11.60	445	6.78	440	3.15	185	33.4	185	25.7	185	22.2
L1Min-PCA	280	4.68	340	3.54	180	2.34	145	15.0	185	9.20	165	6.90



(a) (b) (c) (d)

Fig. 2. Bases learned for *ORL* by different methods. (a) L2PCA; (b) L1PCA_G; (c) L1PCA_NG; and (d) L1Min-PCA.

with 9 poses and 64 illustration conditions. The third database is ORL[3], in which there are ten different images for each of 40 distinct subjects. All the images in this database were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). All the face images have been manually aligned and cropped. The size of each cropped image is 32×32 pixels with 256 gray levels per pixel. Therefore, a 1024-dimensional vector represents an image. Besides face databases, we also used COIL100 database[4] for performance evaluation, which contains images of 100 objects. For each object, there are 72 images taken 5 degrees apart as the object is rotated on a turning table. Some samples from testing database are shown in Fig.1.

[3]http://www.cad.zju.edu.cn/home/dengcai/
[4]http://www.cad.zju.edu.cn/home/dengcai/

A. Projection Bases

First, we present the learned bases, i.e. $\mathbf{P}$, by using different PCA methods before evaluating the performance of L1Min-PCA on supervised and unsupervised learning tasks. Figures 2 and 3 plot the bases estimated for data sets 'ORL' and 'COIL100', respectively. As we can see from Figures 2 and 3, the proposed L1Min-PCA can find the major components of the data set as other methods. There is no essential visual difference among those bases sets obtained by different methods. However, as we shall see later, sparsity associated with L1Min-PCA has advantage in subsequent tasks.

B. Unsupervised Learning

We evaluate the performance and accuracy of face clustering using K-Means on the projected data obtained by different methods. As a benchmark, K-means [8] is directly applied to the original data for comparison. In general, the clustering result is assessed by the number of misclassified samples when the ground truth is available. In this paper, we apply a criterion,

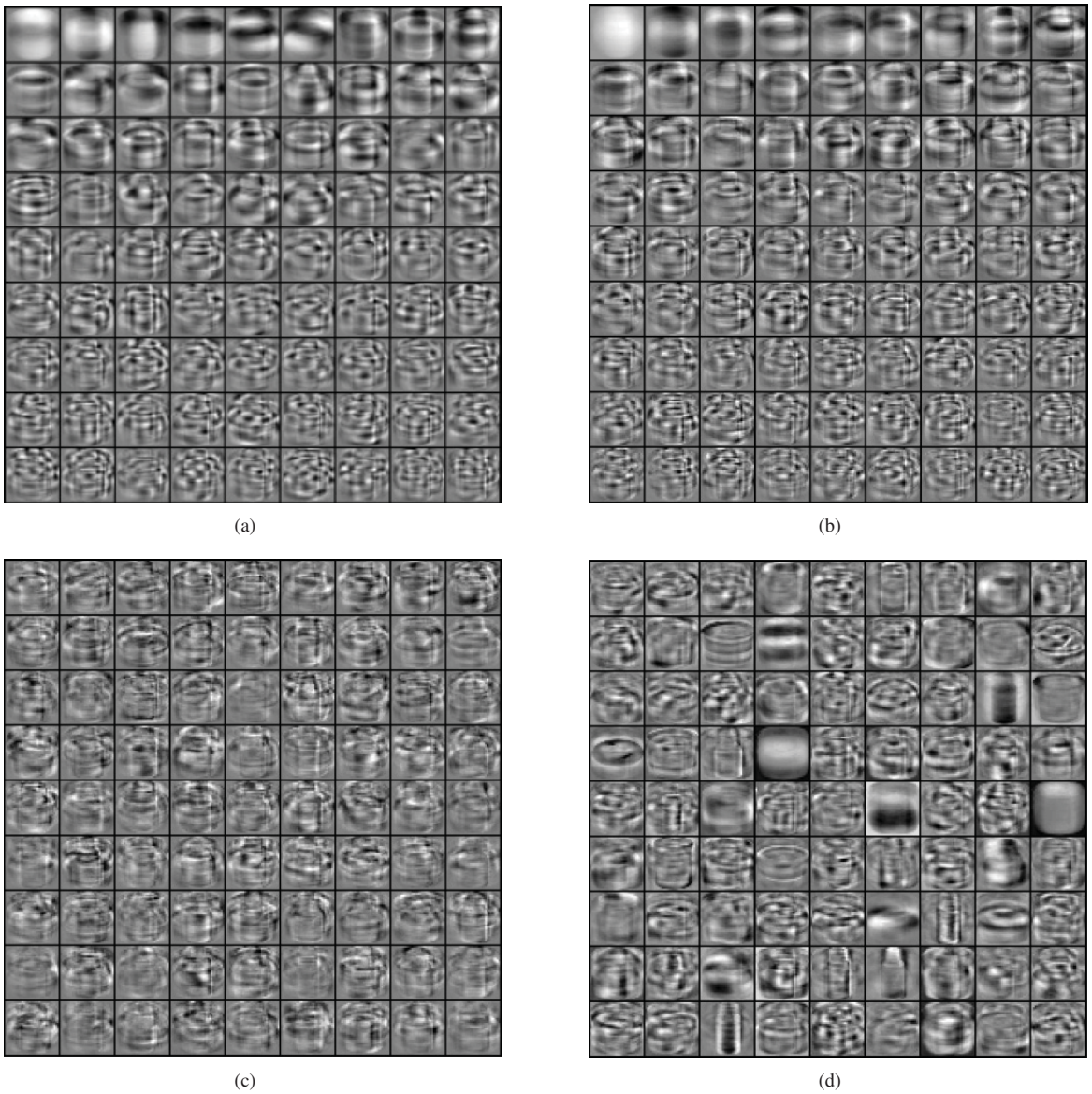


Fig. 3. Bases learned for *COIL100* with different methods (from top to bottom: (a) L2PCA; (b) L1PCA_G; (c) L1PCA_NG; and (d) L1Min-PCA.)

termed Normalized Mutual Information (NMI) [4], to quantify the clustering performance more precisely.

Let K be the set of ground truth clusters and K' the clusters obtained by a clustering algorithm. The mutual information metric $MI(K, K')$ is defined as

$$\mathrm{MI(K, K')} = \sum_{\mathrm{k_i} \in K, k'_j \in K'} p\left(k_i, k'_j\right) \log_2 \frac{p\left(k_i, k'_j\right)}{p\left(k_i\right) p\left(k'_j\right)},$$

where $p(c)$ is the probability of a sample belonging to cluster c and $p(c_1, c_2)$ the probability of a sample from both cluster c_1 and c_2. Then, the NMI is represented by

$$\mathrm{NMI(K, K')} = \frac{\mathrm{MI(K, K')}}{\max\left(\mathrm{H\left(K\right), H\left(K'\right)}\right)},$$

where $\mathrm{H\left(K\right)}$ and $\mathrm{H\left(K'\right)}$ denote the entropy of K and K', respectively. Usually, $\mathrm{NMI(K, K')}$ ranges from 0 to 1. 1 means the two sets of clusters are identical and 0 means that the two are independent.

Figure 4(a) shows the quality of different clustering solutions on PIE database with different target dimensions evaluated by NMI. K is set as the ground truth cluster of the data. All methods can achieve stable performance with dimensionality ranging from 40 to 90. However, when dimensionality varies from 5 to 150, K-Means with projected data obtained by L1Min-PCA is consistently better than K-Means with original data and reduced data obtained by other methods. This suggests that L1Min-PCA can find a subspace in which the data show better cluster structure than in the original space, even when the dimensionality is as low as 5. From clustering point of view, dimensionality reduction by L1Min-

PCA reduces redundant and possible misleading information successfully and therefore improves the clustering result. In contrast, other methods failed to find suitable subspaces for clustering. Although they managed to have comparable performance with baseline method when dimensionality is more than 40, the situation is certainly worse when dimensionality drops below 20.

We repeated the same experiment on **Yale-B** database. The results are shown in Figure 4(b). Similar observations can be made from the figure. These experiments demonstrate that the proposed L1Min-PCA is capable of capturing cluster structure in low dimensional space for test databases.

C. Supervised Learning

Face recognition is a typical supervised learning task [11]. For PIE face database, we randomly selected $r = 33, 50, 66\%$ samples for training. The rest of the samples were used for testing. The recognition was carried out by using the nearest neighbor classifier on the subspace learned in training (That is, $K = 1$). Precisely, the testing images were projected onto lower dimensional subspace by using the bases learned from training images, and then the recognition was performed. For fair comparison, we average the recognition results given for given value of r over 10 repeats.

We evaluated the recognition error rate versus dimensionality on **PIE** and **YaleB** databases. Figure 5 shows the results of different methods for this task. Overall, L1Min-PCA outperforms other methods in terms of classification error rates. Especially when the size of training set is small, L1Min-PCA is better than others by a large margin. Although it is not much so when more data are available for training on **PIE** database, L1Min-PCA leads quite a lot on **YaleB** database. This once again confirms the capability of L1Mn-PCA in finding the meaningful subspace suitable for clustering and classification. Another advantage of L1Min-PCA is that it produces sparse embeddings of the data (the projected data in lower dimensional subspace), which is known to be more computational efficient for subsequent tasks such as classification.

In Table I, we report the best recognition results that each method can possibly achieve by repeating the same experiment with different target dimensionality. Interestingly, for L1Min-PCA, it requires less dimensionality than other methods to achieve better classification rate for both datasets. This result shows that seeking a sparse representation of the data in lower dimensional space is promising in improving the classification rate.

VI. Conclusion

In this paper, a novel linear subspace learning method called L1 Norm Minimization PCA (L1Min-PCA) has been proposed. It minimizes L1-norm of the dispersion in low dimensional feature space so that an optimal subspace with maximum sparsity can be achieved. To avoid greedy search and smooth approximation to L1 norm, we relaxed the orthonormality condition of the projection, and the optimization is carried out by ADM. The efficacy of our proposed L1Min-PCA is verified using clustering and classification experiments on faces images. The quantitative analysis of the experimental results indicates that L1Min-PCA outperforms other similar PCA methods. It is able to reduce the redundant information and recover the cluster structure of data in low dimensional space.

Acknowledgments

Ming Yin's work is supported by the Foundation of Key Laboratory of Autonomous Systems and Networked Control, Ministry of Education, P.R. China (No. 2013A06). Junbin Gao's work is supported by the Australian Research Council (ARC) through Discovery Project Grant DP140102270.

References

[1] A Beck and M Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[2] M. Belkin and P. Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. *Journal of Comput. Syst. Sci.*, 74(8):1289–1308, 2008.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, volume 3. Foundations and Trends in Machine Learning, 2011.

[4] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression: A unified approach for sparse subspace learning. In *ICDM*, pages 73–82. IEEE Computer Society, 2007.

[5] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(1):4–13, Jan. 2005.

[6] F. de la Torre and M. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1):117 –142, 2003.

[7] C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: rotational invariant l1-norm principal component analysis for fobust subspace factorization. In *in Proc. International Conference on Machine Learning*, Pittsburgh, PA, June 2006.

[8] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04. ACM, 2004.

[9] Dingcheng Feng, Feng Chen, and Wenli Xu. Learning robust principal components from l1-norm maximization. *Journal of Zhejiang University - Science C*, 13(12):901–908, 2012.

[10] Junbin Gao. Robust l1 principal component analysis and its bayesian variational inference. *Neural Computation*, 20(2):555–572, 2008.

[11] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong J. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[12] Wei Hong, John Wright, Kun Huang, and Yi Ma. Multi-scale hybrid linear models for lossy image representation. *IEEE Trans. on Image Processing*, 12:3655–3671, 2006.

[13] M J. Ho, H. Yang, J. Lim, K.C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

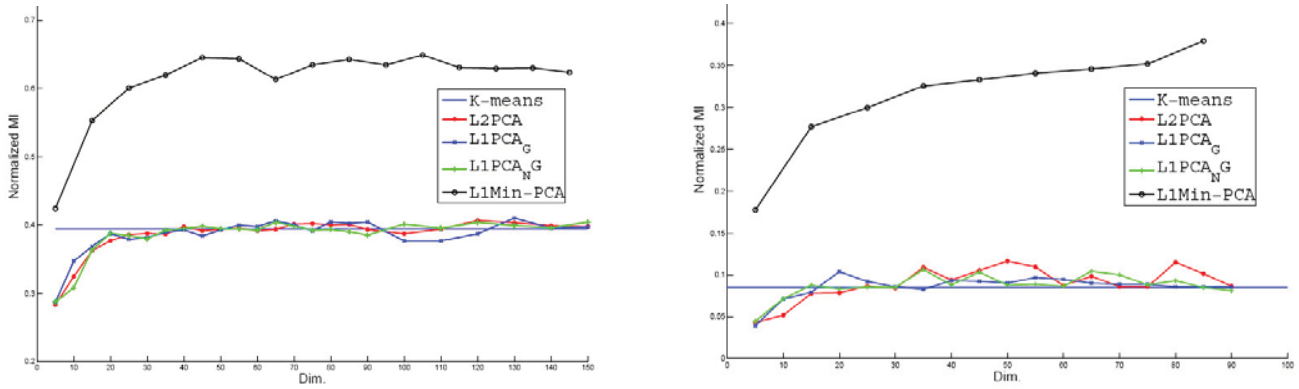
[14] Xudong Jiang. Linear subspace learning-based dimensionality reduction. *IEEE Signal Process. Mag.*, 28(2):16–26, 2011.

[15] Q. Ke and T. Kanade. Robust *l*1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2005.

[16] N. Kwak. Principal component analysis based on l1-norm maximization. *IEEE Transactions on PAMI*, 30(9):1672–1680, 2008.

[17] F. De la Torre and M.J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1/2/3):117–142, August 2003.

[18] Q.V. Le, A. Karpenko, J. Ngiam, and A.Y. Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 1017–1025, 2011.



(a) (b)

Fig. 4. Normalized mutual information vs dimensionality, (a) *PIE*, (b) *Yale-B*.

[19] Qiao Lishan, Chen Songcan, and Tan Xiaoyang. Sparsity preserving projections with applications to face recognition. *Pattern Recogn.*, 43(1):331–341, January 2010.

[20] Feiping Nie, Heng Huang, Chris Ding, Dijun Luo, and Hua Wang. Robust principal component analysis with non-greedy $l1$-norm maximization. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pages 1433–1438. AAAI Press, 2011.

[21] Bruno A. Olshausen and David J. Fieldt. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997.

[22] R. Vidal, R. Tron, and R. Hartley. Multiframe motion segmentation with missing data using powerfactorization and gpca. *International Journal of Computer Vision*, 79:85–105, 2008.

[23] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(1):40–51, 2007.

[24] Allen Y. Yang, John Wright, Yi Ma, and Shankar Sastry. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2):212–225, 2008.

[25] Lei Zhang, Pengfei Zhu, Qinghua Hu, and Zhang David. A linear subspace learning approach via sparse coding. In *ICCV*, pages 755–761. IEEE, 2011.

[26] H. Zhou, T. Hastie, and R. Tibshirani. Sparse principle component analysis. Technical report, Statistics Department, Stanford University, 2004.

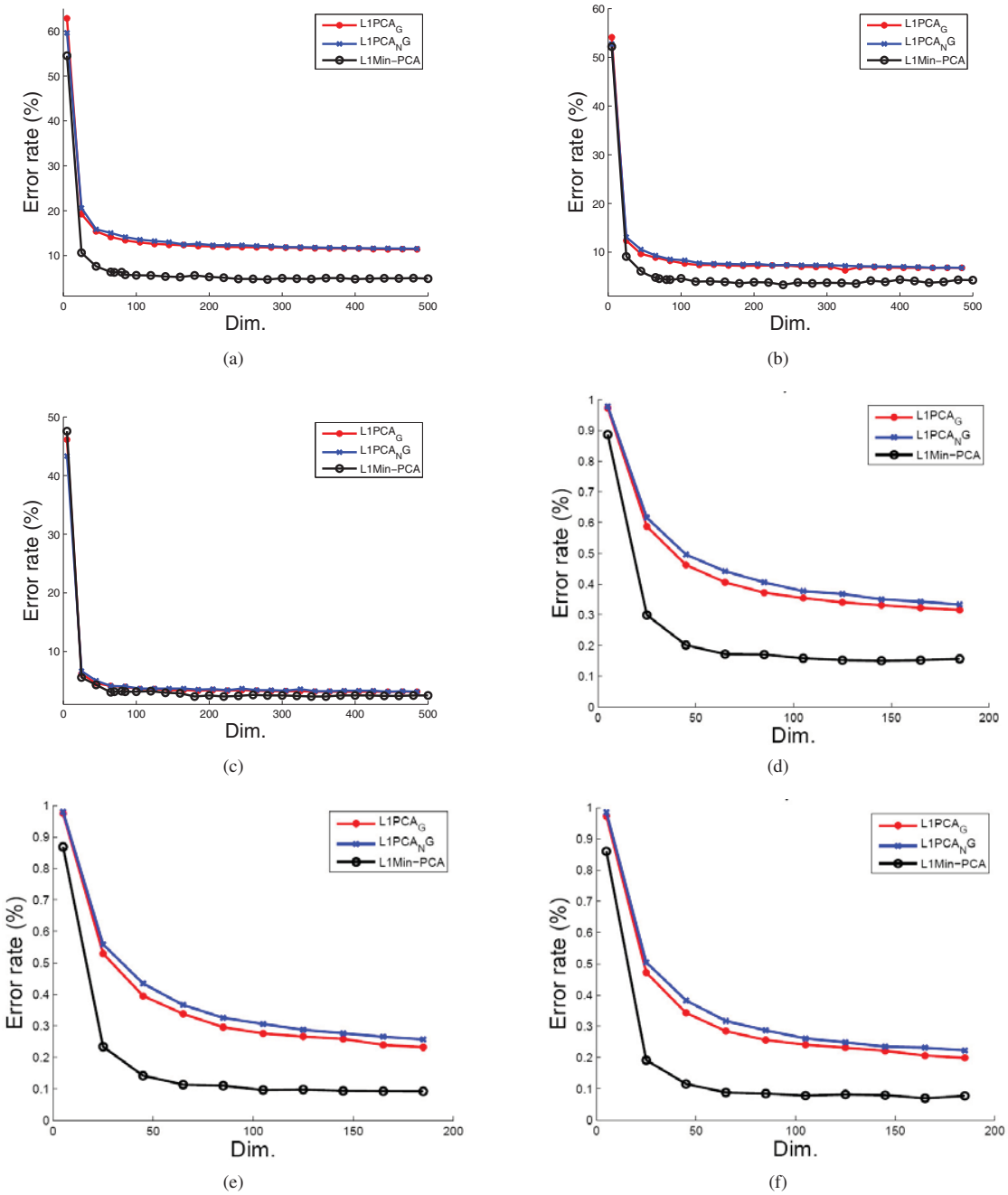


Fig. 5. Error rate vs dimensionality reduction on *PIE* (the first row) and *Yale-B* (the second row) databases: (a) and (d) 33 % training; (b) and (e) 50 % training; (c) and (f) 66 % training.