STRUCTURED SPARSE LINEAR DISCRIMINANT ANALYSIS

Zhen Cui[1,2], *Shiguang Shan*[1], *Haihong Zhang*[3], *Shihong Lao*[3], *and Xilin Chen*[1]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]Scholl of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[3]Omron Social Solutions Co., LTD., Kyoto, Japan
{zhen.cui, shiguang.shan}@vipl.ict.ac.cn; angelazhang@ssb.kusatsu.omron.co.jp; lao@ari.ncl.omron.co.jp; xlchen@ict.ac.cn

ABSTRACT

Linear Discriminant Analysis (LDA) is an efficient image feature extraction technique by supervised dimensionality reduction. In this paper, we extend LDA to Structured Sparse LDA (SSLDA), where the projecting vectors are not only constrained to sparsity but also structured with a pre-specified set of shapes. While the sparse priors deal with small sample size problem, the proposed structure regularization can also encode higher-order information with better interpretability. We also propose a simple and efficient optimization algorithm to solve the proposed optimization problem. Experiments on face images show the benefits of the proposed structured sparse LDA on both classification accuracy and interpretability.

Index Terms— Linear discriminant analysis, Sparse coding, Face recognition, Least squares, Interpretability

1. INTRODUCTION

Linear Discriminant Analysis (LDA) is an efficient method for image feature extraction and dimensionality reduction. Its goal is to separate the class means well and simultaneously achieve small within-class variance via projecting high-dimensional data into a low-dimensional space.

One of LDA's shortcomings is that, even though it can find a small number of important projections, the mappings typically involve all original variables. Therefore, when the number of variables is large, the within-class covariance matrix is hard to be reliably estimated, and thus overfitting often occurs. To avoid the overfitting on high dimension data and small sample size, in recent years, some variations of LDA which find sparse factors have been developed. For example, Moghaddam et al. [1] proposed an exact and optimal sparse LDA (SLDA) using spectral branch-and-bound search with high complexity. Qiao et al. [2] also proposed an efficient algorithm to deal with the overfitting phenomenon by resorting to the connection of Fisher's LDA [3] and generalized eigenvalue problem. In fact, the sparse LDA essentially implies variable selection. For face images, the sparsity can capture some features or parts of face which might be intuitively meaningful from our prior. More generally, it is desirable to learn higher-order information which reflects the structure of the face so that we may enforce certain constraints on face components to promote recognition performance.

In recent years, some researches [4,5,6,7] on structured sparsity have demonstrated the benefit from such structured prior in the context of compressed sensing, regression, classification, and so on. In particular, Jenatton et.al [7] proposed the structured sparse principal component analysis (SSPCA), where the sparse patterns of all dictionary atoms are structured and constrained to be a pre-specified groups by exploiting $l_{2,1}$ mixed norm. SSPCA gives a good explanation on which area is more important and meaningful for face recognition.

Motivated by the above analysis, in this paper, we aim to go beyond sparse LDA and propose structured sparse LDA (SSLDA). Different from previous LDA and its variations, SSLDA not only models the data with well separability, but also mines some priori structure information to explain the variance implied in the data. We demonstrate that Least Squares LDA (LSLDA) [8] can be used successfully to establish the model with structured sparsity-inducing norms. We also propose a simple and efficient optimization algorithm to solve this model. Experimental results on face images show the benefits of SSLDA on recognition performance and interpretability.

2. LINEAR DISTCRIMINANT ANALYSIS

Consider the input data vectors $X = (\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots, \boldsymbol{x}_n) \in R^{p \times n}$ in high dimensional space, where p is the original feature dimension and n is the sample number. For simplicity, the data is centered so that $\bar{\boldsymbol{x}} = \sum_i \boldsymbol{x}_i / n = 0$. In LDA, we use the between-class scatter and within-class scatter matrices:

$$S_b = \sum_k n_k \boldsymbol{m}_k \boldsymbol{m}_k^T, S_w = \sum_k \sum_{i \in C_k} (\boldsymbol{x}_i - \boldsymbol{m}_k)(\boldsymbol{x}_i - \boldsymbol{m}_k)^T \quad (1)$$

where $\boldsymbol{m}_k$, n_k is the center and sample number of the *k*-th cluster respectively. The LDA's projections *V* are determined by

$$\max_V \frac{trace(V^T S_b V)}{trace(V^T S_w V)} , \quad (2)$$

which can be solved as a generalized eigenvalue equation,

$$S_b \boldsymbol{v} = \lambda S_w \boldsymbol{v} \quad (3)$$

with the leading eigenvectors.

The above LDA formulation has been proven to be equivalent to a least squares problem under a mild condition, i.e. least squares LDA (LSLDA) [8], that is,

$$\min_V \left\| Y - X^T V \right\|, \quad (4)$$

where $Y \in R^{n \times c}$, and *c* is the cluster number, and *Y* is defined by

$$Y(i,j) = \begin{cases} \sqrt{n/n_j} - \sqrt{n_j/n} & \text{if } y_i = j, \\ -\sqrt{n_j/n} & \text{otherwise.} \end{cases} \quad (5)$$

where n_j is the number of samples in the *j*-th cluster. It has been proven theoretically and empirically that the equivalence between LSLDA (4) and original LDA (2) holds for most high-dimensional and under-sampled data [8].

3. PROJECTION VIA STRUCTURE SPARSE REGRESSION

In this section, we first introduce structured sparsity-inducing norms, then formulate structured sparse LDA, and finally give an efficient algorithm.

3.1. Structured Sparsity-Inducing Norms

Jenatton et al. [7] introduced a structured sparse norm, where the regularization (e.g. $l_{2,1}$-norm) is exploited to enforce between-group sparsity with within-group l_2 norm. That is, such mixed norm highlights more the selection of a few groups of variables rather than single variable. Generally, Interesting groups are pre-defined/auto-learned by certain priors. Then, following by the pre-specified groups, the structured sparse norm can be formulated as follows,

$$\Omega^\alpha(\boldsymbol{v}) = \left(\sum_{g \in G} \left\| \boldsymbol{d}^g \circ \boldsymbol{v} \right\|_2^\alpha \right)^{1/\alpha} = \left\| \left(\left\| \boldsymbol{d}^g \circ \boldsymbol{v} \right\|_2 \right)_{g \in G} \right\|_\alpha , \quad (6)$$

where *g* is a subset of variable indexes so that $\cup_{g \in G} = \{1, \cdots, p\}$ covers all variables, $(\boldsymbol{d}^g)_{g \in G} \in R^{p \times |G|}$ is a $|G|$ -tuple of *p*-dimensional vectors such that $d_i^g > 0$ if $i \in g$ and 0 otherwise, $\alpha \in (0,2)$, the operation "$\circ$" means the dot product. Especially, when $\alpha = 1$, i.e. the l_1 norm is used to constraint to group selection, thus the above structured constraint (6) is $l_{2,1}$ norm which linearly combines the l_2 norms of possibly groups of variables with the weight $\boldsymbol{d}^g$. In practice, we exploit the l_1 norm since it is convex.

Another question is how to define the groups. There are many options, in this paper, but we merely investigate the form of 2-dimensional grid on images. Following by the strategy in [7], we partition the 2-dimensional plane in all horizontal and vertical half-spaces with overlapping groups, which has been proven to be able to induce rectangular nonzero patterns under l_1 norm.

3.2. Formulation of Structured Sparse LDA

By combining (4) with (6), we can obtain the model of structured sparse LDA,

$$\arg\min_V \ \frac{1}{2c} \left\| Y - X^T V \right\|_F^2 + \lambda \sum_{k=1}^{c} \Omega^\alpha(V^k) , \quad (7)$$

where V^k is the *k*-th column of matrix *V*. Since each column of *V* is independent in this regression model, we may operate on each column on *Y*. Therefore, (7) can be converted into the following formulation for simplicity,

$$\arg\min_{\mathrm{v}} \frac{1}{2c} \left\| \mathrm{y} - X^T \mathrm{v} \right\|_F^2 + \lambda \Omega^\alpha(\mathrm{v}) . \quad (8)$$

However, this optimization problem is no longer differentiable and convex when $\alpha < 1$. To address this regularization, we utilize the following theorem from [7].

Theorem 1: Let $\alpha \in (0,2)$ and $\beta = \alpha/(2-\alpha)$. For any vector $\boldsymbol{u} \in R^m$, we have the following equality,

$$\left\| \boldsymbol{u} \right\|_\alpha = \min_{\boldsymbol{u} \in R_+^m} \frac{1}{2} \sum_{i=1}^{m} \frac{u_j^2}{z_j} + \frac{1}{2} \left\| z \right\|_\beta ,$$

and the minimum is uniquely attained for $z_j = \left| u_j \right|^{2-\alpha} \left\| \boldsymbol{u} \right\|_\alpha^{\alpha-1}, \forall j \in \{1, \cdots m\}$.

Using above theorem, we can reformulate the objective function (8) as follows,

$$\arg\min_{\mathrm{v},\mathrm{z}} \ f(\mathrm{v},\mathrm{z}) = \frac{1}{2c} \left\| \mathrm{y} - X^T \mathrm{v} \right\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^{|G|} \left\| \mathrm{d}^{g_i} \circ \mathrm{v} \right\|_2^2 / z_i + \frac{\lambda}{2} \left\| \mathrm{z} \right\|_{\alpha/(2-\alpha)} . \quad (9)$$

3.3. Optimization

To solve the optimization problem in (9), we present an algorithm alternating the calculation of the two variables involved, i.e., $\boldsymbol{v}$ and $\boldsymbol{z}$. So, the optimization algorithm mainly includes two steps: first, we use Theorem 1 to solve $\boldsymbol{z}$ in (9) for fixed $\boldsymbol{v}$. Second, for fixed $\boldsymbol{z}$, we can solve $\boldsymbol{v}$ as the following convex problem,

$$f(\mathrm{v}) = \frac{1}{2c} \left\| \mathrm{y} - X^T \mathrm{v} \right\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^{|G|} \left\| \mathrm{d}^{g_i} \circ \mathrm{v} \right\|_2^2 / z_i = \frac{1}{2c} \left\| \mathrm{y} - X^T \mathrm{v} \right\|_F^2 + \frac{\lambda}{2} \mathrm{v}^T \Lambda \mathrm{v} , \quad (10)$$

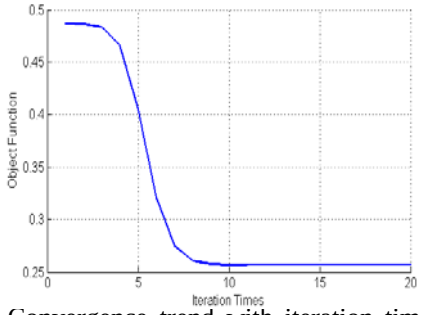

Figure 1. Convergence trend with iteration times. Data comes from AR [9] database.

where $\Lambda = diag(\sum_{i=1}^{|G|}((\boldsymbol{d}^{g_i} \circ \boldsymbol{d}^{g_i})/z_i))$. Thus (10) has a closed form solution for the variable $\boldsymbol{v}$, which can be updated as

$$\mathrm{v} = \left(XX^T + c\lambda\Lambda\right)^{-1} X\mathrm{y} \qquad (11)$$

By alternating the two steps, our algorithm can converge very fast, as shown in Fig.1. In experiments, the stopping criterion relies on the relative decrease (e.g. 10^{-3}) in the cost function (8). More importantly, the algorithm does not need a warm starting.

4. EXPERIMENTS

We use two standard face databases, AR [9] and Extended YaleB [10], to experimentally compare the proposed SSLDA method with other methods by the nearest neighbor classifier. The comparison methods include Regularized LDA (RLDA) [11], Uncorrelated LDA (ULDA) [12], Orthogonal LDA (OLDA) [12], LSLDA and SLDA. For fair comparison, SLDA is implemented with only sparsity instead of group sparsity based on the regression model of LDA. For clarity, we only consider L1 sparse, i.e. α=1, and present results for SSLDA and SLDA with the parameters λ=0.005, 0.1 unless specified. The projecting dimension of LDA is set to class number minus 1. The $\boldsymbol{z}$ is initialized as $\mathbf{0}$.

A subset from the AR database [9] consists of 1,400 non-occluded face images, corresponding to 100 subjects (50 males and 50 females), is used in our experiments. We crop and normalize the original face images to 40*32 pixels by aligning two eyes. The Extended Yale B [10] contains about 2,414 frontal face images of 38 individuals with varying illumination condition. We also use the cropped face images of size 40*32 pixels. To eliminate illumination effect partly, we exploit histogram equalization.

We first tested the different algorithms on AR database. We randomly split the subset into two halves. One half, containing 7 face images for each person, is used for training, and the other half is used for testing. Fig. 2 shows some examples of learned projections for SLDA and SSLDA, where the parameter λ is respectively set to 0.02 and 1.0 for more clear description. Each pixel value implies the weight for corresponding location in face image. Positive and

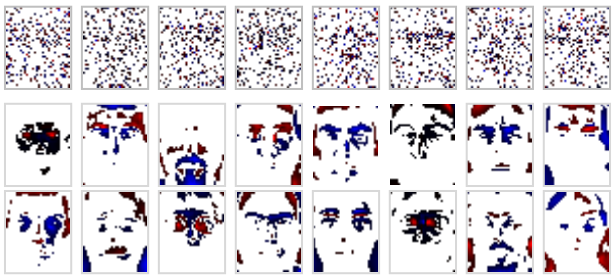

Figure 2. The learned projections on AR database: SLDA (the first row) and SSLDA (the bottom two rows). Each image represents a projecting vector, only a few vectors are shown here because of space limitation. Blue pixels mean negative values, red pixels represent positive values. (Please see the color figure)

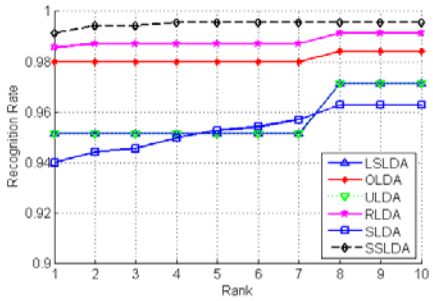

Figure 3. Classification accuracy versus variants of LDA with k-NN classifier on AR database. SSLDA does not degrade the performance of face recognition.

negative values are respectively represented by red and blue color. For more clear representation, we reduce the values close to 0 (less than 10^{-3} with absolute value) to 0 for projections of SSLDA. While SLDA finds sparse but spatially unconstrained patterns, SSLDA chooses sparsely convex areas that approximately correspond to natural components of face image.

We also quantitatively compare SSLDA with other methods, shown in Fig.3. Surprisingly, we find that the performance of SSLDA on the test set is pretty excellent compared with SLDA, which yields sparsity without any structured priors. SSLDA slightly outperforms the RLDA, which improves over ULDA and OLDA. This also implies the importance of the structure in face recognition, which helps SSLDA to improve the recognition performance.

In the second experiment we test the proposed algorithm's ability on Extended Yale B database [10] which consists of face images with complex illumination variations. We randomly choose 4, 8, 16, 32 face images from per subject as the training set, the others images are used to test respectively. The experimental results are shown in Fig. 4. SSLDA outperforms other methods consistently. In particular, for small sample size, SSLDA is more robust. When the number of training samples is 4 and 8, the performance of SSLDA is significantly superior to other best methods. More importantly, SSLDA completely outperforms

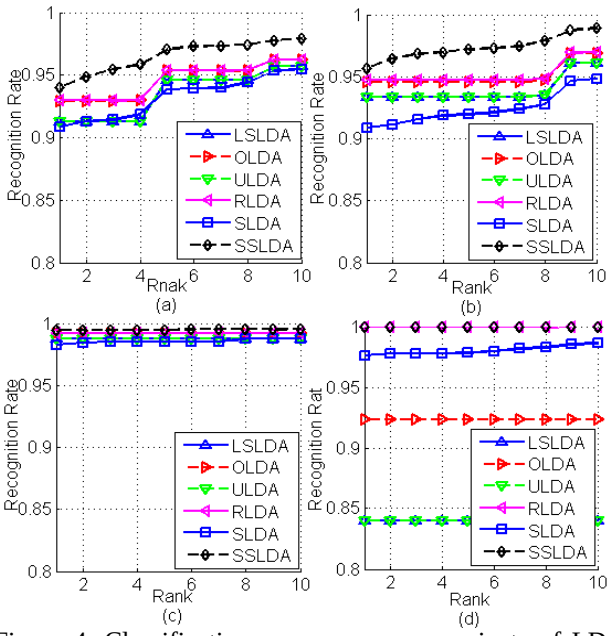

Figure 4. Classification accuracy versus variants of LDA with k-NN classifier on Extended Yale B database. (a), (b), (c) and (d) use 4, 8, 16 and 32 face images per subject as the training set respectively. When training set size is small, the performance of SSLDA significantly outperforms other methods. With increasing size, the RLDA is close to SSLDA on classification accuracy. More importantly, SSLDA improves over SLDA consistently, which implies facial structures help face recognition.

LDA with only sparsity constraints, which again verifies that implicitly facial structure can help face recognition. With increasing sample size, the performance of RLDA is close to SSLDA since enough samples may make learned projection more robust to generalize on testing samples. In addition, we also find the recognition performance becomes better with increasing of the training samples number because the overfitting can be generally reduced when utilizing a large number of training samples.

5. CONCLUSION

In this paper, we proposed to apply the structured sparse to Linear Discriminant Analysis from the viewpoint of reconstruction based on the fact that multivariate LDA is equivalent to a least squares model under a mild condition. We added the sparsity-inducing norms into the least square LDA model and then proposed a simple but efficient algorithm to solve SSLDA. In the task of face recognition, our approach leads to not only better performance but also more interpretable projections.

In future work, we are going to extend the structured sparsity-inducing regularization to linear graph embedding, and explore more applications in object classification.

6. ACKNOWLEDGEMENT

This paper is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; Natural Science Foundation of China under contracts Nos. 61025010, 61173065, 60872124, and 60832004; and Beijing Natural Science Foundation (New Technologies and Methods in Intelligent Video Surveillance for Public Security) under contract No.4111003. Haihong Zhang and Shihong Lao are partially supported by "R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society", Special Coordination Fund for Promoting Science and Technology of MEXT, the Japanese Government.

7. REFERENCES

[1] B. Moghaddam, Y. Weiss and S. Avidan, "Generalized spectral bounds for sparse LDA," in *Proc. ICML*, 2006.

[2] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data," *IAENG Int. J. Applied Math.*, vol, 39, no. 1, 48–60, Feb. 2009.

[3] R.A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Egeniscs*, 7:179-188. 1936.

[4] R.G.Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, 56(4): 1982-2001, 2010.

[5] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proc. ICML*, 2009.

[6] L. Jacob, G. Obozinski, and J. P. Vert, "Group Lasso with overlap and graph Lasso," in *Proc. ICML*, 2009.

[7] R. Jenatton, G. Obozinski, and F. Bach. "Structured sparse principal component analysis," In the *13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[8] J. Ye, "Least squares linear discriminant analysis," in *Proc. ICML*, 2007.

[9] A. Martinez, and R. benavente, "The AR face database," *CVC Tech. Report* , No. 24, 1998.

[10] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE PAMI*, 27(5):684–698, 2005.

[11] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[12] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research*, 6:483-502, 2005.