Sparse approximation to discriminant projection learning and application to image classification

Yu et al., 2019, *Pattern Recognition*

- Topic
  - Subspace learning for classification: $Z = WX \ (m \times n) = (m \times d) \times (d \times n)$
  - LDA-based (discriminant projection): within-class covariance S_W, between-class covariance S_B

- Problems
  - Absence of a sparse solution
  - HDLSS (high-dimensional, low sample size) problem

- Motivation
  - Embedding a feature selection framework into subspace learning
  - A new optimization formulation to avoid singularity problem

- Proposed method: SADPL
  (Sparse Approximation to Discriminant Projection Learning)
  - [Proposition 1] (Gaynanova et al., 2016) There exists a low-rank matrix A such that $S_B = AA^T$ and k-th column of A is

$$A_k = \frac{\sqrt{n_{k+1}}\left(\sum_{r=1}^{k} n_r\left(u^r - u^{k+1}\right)\right)}{\sqrt{n}\sqrt{\sum_{r=1}^{k} n_r \sum_{r=1}^{k+1} n_r}}$$

  where $k = 1, \cdots, C-1$
  - [Proposition 2] There exists an orthogonal matrix Q such that $(S_W + S_B)^{-1}AQ$ is the eigenvectors-matrix of $S_W^{-1}S_B$
  - [Proposition 3] For any orthogonal matrix Q,

$$\|P^T y - P^T x\|_2^2 = (y-x)^T P P^T (y-x)$$
$$= (y-x)^T P Q Q^T P^T (y-x) = \|(PQ)^T y - (PQ)^T x\|_2^2$$

  where P is a projection matrix $P \in \mathbb{R}^{d \times (C-1)}$
  - Error function

$$\min_P \frac{1}{2}\left\|(S_W + S_B)^{1/2}P - (S_W + S_B)^{-1/2}A\right\|_F^2$$
$$= \min_P \frac{1}{2} Tr(P^T S_W P) + \frac{1}{2}\|A^T P - I\|_F^2$$

- F-norm regularizer

$$\|P\|_F = \sqrt{\sum_{i=1}^{d}\|P^i\|_2^2} = \sqrt{\sum_{j=1}^{C-1}\left\|P_j\right\|_2^2}$$

  where P^i and P_j are i-th row and j-th column of P, respectively
- $L_{2,1}$-norm regularizer

$$\|P\|_{2,1} = \sum_{i=1}^{d}\left\|P^i\right\|_2 = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{C-1} P_{ij}^2}$$

- Convex optimization problem

$$\underset{P}{\operatorname{argmin}} J(P) = \frac{1}{2}Tr(P^T S_W P) + \frac{1}{2}\|A^T P - I\|_F^2 + \frac{\lambda_1}{2}\|P\|_F^2 + \lambda_2\|P\|_{2,1}$$
$$\frac{\partial J(P)}{\partial P} = S_W P + A(A^T P - I) + \lambda_1 P + \lambda_2 BP$$
$$P = (S_W + AA^T + \lambda_1 I + \lambda_2 B)^{-1}A$$

  where B is a diagonal matrix with $B_{i,i} = \frac{1}{2\|P^i\|_2}$
- Since B is dependent of P,

Algorithm 1 SADPL algorithm.

Input: $\mathbf{X} \in R^{d\times n}$, λ_1 and λ_2.
Output: $\mathbf{P} \in R^{d\times C-1}$.
1: Compute within-class scatter matrix $\mathbf{S_W}$ and between-class scatter matrix $\mathbf{S_B}$;
2: Compute the matrix $\mathbf{A}$ based on Proposition 1;
3: Initialize $t = 0$;
4: Initialize $\mathbf{B_0}$;
5: **while** Not convergent **do**
6: $\quad \mathbf{P}^{[t+1]} = (\mathbf{S_W} + \mathbf{AA}^T + \lambda_1\mathbf{I} + \lambda_2\mathbf{B}^{[t]})^{-1}\mathbf{A}$;
7: $\quad$ Update $\mathbf{B}^{[t+1]}$, here the ith diagonal element $B_{i,i}^{[t+1]} = 1/2\|(\mathbf{p}^{[t+1]})^i\|_2$;
8: $\quad t = t + 1$;
9: **end while**

- [Theorem 1] In Algorithm 1, the value of $J(P)$ monotically decreases along with the iteration

➢ Experiments
- 5 image databases
- 1NN classification after projection
- Sample results

Table 3
Classification accuracies for different approaches on the KTH-TIPS texture database (Mean ± STD (%)).

Method	$T_s = 20$	$T_s = 30$	$T_s = 40$
PCA	75.33 ± 3.27	79.92 ± 2.62	82.88 ± 2.71
LDA	80.83 ± 1.70	84.83 ± 2.69	87.47 ± 2.59
LPP	81.92 ± 2.24	86.08 ± 2.32	88.08 ± 2.11
LSDA	45.67 ± 7.59	42.12 ± 2.12	33.08 ± 11.82
SRDA	82.78 ± 1.67	86.25 ± 2.43	88.14 ± 2.36
RILDA	48.43 ± 14.33	57.33 ± 13.31	58.75 ± 11.14
SRRS	75.91 ± 2.98	80.92 ± 2.74	84.57 ± 2.66
LLDA	75.80 ± 3.15	79.87 ± 2.57	85.61 ± 2.71
SADPL	**84.37 ± 2.36**	**88.62 ± 2.15**	**90.41 ± 2.26**

- Effect of regularization terms

Table 6
Classification accuracies of SADPL under different regularization terms (%).

Database	FRGC	2D+3D	KTH-TIPS	COIL-20	CIFAR-10
Adding F-norm	99.33	99.28	88.62	99.32	91.25
Without F-norm	99.07	96.57	86.37	93.14	91.08
Using l_1-norm	99.27	99.28	88.59	99.31	91.20

➢ References

- Gaynanova et al., 2016, Simultaneous sparse estimation of canonical vectors in the $p \gg n$ setting (Proposition 1)
- Turk and Pentland, 1991, Eigenfaces for recognition (PCA)
- Belhumeur et al., 1997, Eigenfaces vs. fisherfaces: recognition using class specific linear projection (LDA)
- Cai et al., 2008, SRDA: an efficient algorithm for large-scale discriminant analysis (SRDA)
- Lai et al., 2017, Rotational invariant dimensionality reduction algorithms (RILDA)
- Cai et al., 2007, Locality sensitive discriminant analysis (LSDA)
- He and Niyogi, 2003, Locality preserving projections (LPP)
- Fan et al., 2011, Local linear discriminant analysis framework using sample neighbors (LLDA)
- Li and Fu, 2016, Learning robust and discriminative subspace with low-rank constraints (SRRS)