ALIGNMENT WITH INTRA-CLASS STRUCTURE CAN IMPROVE CLASSIFICATION

Jiaji Huang, Qiang Qiu*, Robert Calderbank*, Miguel Rodrigues†, Guillermo Sapiro**

*Department of Electrical Engineering, Duke University
†Department of Electronic & Electrical Engineering, University College London

ABSTRACT

High dimensional data is modeled using low-rank subspaces, and the probability of misclassification is expressed in terms of the principal angles between subspaces. The form taken by this expression motivates the design of a new feature extraction method that enlarges inter-class separation, while preserving intra-class structure. The method can be tuned to emphasize different features shared by members within the same class. Classification performance is compared to that of state-of-the-art methods on synthetic data and on the real face database. The probability of misclassification is decreased when intra-class structure is taken into account.

***Index Terms*—** subspace, principal angle, classification, feature extraction

1. INTRODUCTION

Data that is nominally high dimensional often exhibits a low dimensional geometric structure. For example, frontal images of human faces are recorded by more than 1000 pixels, but can be represented by a 9-dimensional harmonic subspace [1]. Motion trajectories of a rigid body might be recorded by hundreds of sensors, but must intrinsically be represented by a 3-dimensional subspace. There are many more examples where a low-dimensional subspace model captures intrinsic geometric structure and where features extracted from the representation are very effective in typical machine learning tasks such as classification and clustering [2, 3].

Feature extraction and dimension reduction are of great interest to both the signal processing and machine learning communities. There is considerable work in the signal processing community dedicated to the design and analysis of dimension reduction provided by compressive sensing matrices [4, 5]. There has been much interest within machine learning on the discriminative power of linear and non-linear embeddings [3, 6]. The combination of subspace representation with information theoretic analysis [7] has led to a variety of methods for dimension reduction that are able to extract task-specific features [8], for signal reconstruction [9], classification [10, 11], etc.

We consider the problem of discriminating K classes, each of which is approximated by a low-rank linear subspace. When the approximation provided by the subspace model is sufficiently accurate, we show that the error probability of an optimal classifier is determined by the product of the sines of the principal angles between subspaces. This result motivates the design of a new method, TRAIT, that extracts low dimensional discriminative features while preserving intra-class structure. In contrast to a state-of-the-art discriminative feature extraction method, LRT [3], which compresses the data within classes, TRAIT is tunable to preserve user selected intra-class structure. Competitive classification accuracies are achieved with TRAIT features alone on both synthetic and real data. However, classification based on a combination of LRT and TRAIT features is superior to classification based on only one of these approaches.

2. PRINCIPAL ANGLES

Consider two subspaces $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{R}^n$ with dimensions ℓ and s respectively, where $\ell \leq s$. Denote the principal angles between $\mathcal{X}$ and $\mathcal{Y}$ by $\theta^{(1)}, \ldots, \theta^{(\ell)}$, defined recursively as follows

$$\begin{aligned}
\theta^{(1)} &= \min\nolimits_{x_1 \in \mathcal{X}, y_1 \in \mathcal{Y}} \arccos\left(\frac{x_1^\top y_1}{\|x_1\|\|y_1\|}\right), \\
&\vdots \\
\theta^{(j)} &= \min_{\substack{x_j \in \mathcal{X}, y_j \in \mathcal{Y} \\ x_j \perp x_1, \ldots, x_{j-1} \\ y_j \perp y_1, \ldots, y_{j-1}}} \arccos\left(\frac{x_j^\top y_j}{\|x_j\|\|y_j\|}\right), \quad j = 2, \ldots, \ell
\end{aligned}$$

The vectors $x_1, \ldots, x_\ell$ and $y_1, \ldots, y_\ell$, are called principal vectors. The dimension of $\mathcal{X} \cap \mathcal{Y}$ is the multiplicity of zero as a principal angle. It is straightforward to compute the principal angles by calculating the singular values of $X^\top Y$, where X and Y are orthonormal bases for $\mathcal{X}$ and $\mathcal{Y}$ respectively. The singular values of $X^\top Y$ are $\cos\theta^{(1)}, \ldots, \cos\theta^{(\ell)}$.

3. ERROR PROBABILITY OF CLASSIFYING LINEAR SUBSPACES

We model K-class classification as a hypothesis testing problem, where the k-th hypothesis is

$$H_k : x \sim \mathcal{N}(0, \Sigma_k), \quad 1 \leq k \leq K. \tag{1}$$

The covariance matrix Σ_k is assumed to be approximately low-rank and we are interested in the error probability of the Maximum A Posteriori (MAP) classifier, which is optimal.

3.1. Upper Bound of Error Probability

We focus on the case $K = 2$, since the generalization from two to multiple classes has been studied extensively [12, 13]. Denote the signal subspaces of these two classes as $\mathcal{X}$ and $\mathcal{Y}$ respectively. For simplicity we assume that the two hypotheses are equiprobable and that $dim(\mathcal{X}) = dim(\mathcal{Y}) = d$. This restriction is consistent with data sets of faces ($d = 9$) and motion trajectories ($d = 3$). The principal angles (in ascending order) between $\mathcal{X}$ and $\mathcal{Y}$ are $\theta^{(1)}, \ldots, \theta^{(d)}$. Suppose $dim(\mathcal{X} \cap \mathcal{Y}) = r$ so that $\theta^{(1)} = \cdots = \theta^{(r)} = 0$. It follows that the corresponding covariances Σ_1, Σ_2 have the following SVD:

$$\begin{aligned} \Sigma_1 &= U_{1,o}\Lambda_{1,o}U_{1,o}^\top + U_{1,n}\Lambda_{1,n}U_{1,n}^\top + \sigma^2 U_{1,\perp}U_{1,\perp}^\top, \\ \Sigma_2 &= U_{2,o}\Lambda_{2,o}U_{2,o}^\top + U_{2,n}\Lambda_{2,n}U_{2,n}^\top + \sigma^2 U_{2,\perp}U_{2,\perp}^\top, \end{aligned} \tag{2}$$

where both $U_{1,o} \in \mathbb{R}^{n\times r}$ and $U_{2,o} \in \mathbb{R}^{n\times r}$ span $\mathcal{X} \cap \mathcal{Y}$ with singular values $\Lambda_{1,o}$ and $\Lambda_{2,o}$. $U_{1,n} \in \mathbb{R}^{n\times(d-r)}$ spans $\mathcal{X}\backslash\mathcal{Y}$ with singular values $\Lambda_{1,n}$. And $U_{2,n} \in \mathbb{R}^{n\times(d-r)}$ spans $\mathcal{Y}\backslash\mathcal{X}$ with singular values $\Lambda_{2,n}$.

In the next section, we will sometimes use $U_k \triangleq [U_{k,o}, U_{k,n}]$ to denote the signal subspace of class k. Finally $U_{1,\perp}, U_{2,\perp} \in \mathbb{R}^{n\times(n-d)}$ span the noise subspaces with a variance of σ^2. We now characterize the behavior of the classification error in the following theorem:

Theorem 1. *Let $K = 2$ in (1) and let Σ_1, Σ_2 be as defined in (2). As the variance $\sigma^2 \to 0$, the error probability, P_e, of the optimal classifier is upper bounded by*

$$P_e \le \bar{P}_e = c_r(\sigma^2)^{\frac{d-r}{2}} \left(\prod_{j=r+1}^{d} \sin^2\theta^{(j)}\right)^{-\frac{1}{2}} + o(\sigma^{d-r}). \tag{3}$$

The individual signal energies $\Lambda_{1,n}$, $\Lambda_{2,n}$, $\Lambda_{1,o}$, $\Lambda_{2,o}$ and the intersection dimension r determine

$$c_r = 2^{d-\frac{r}{2}} \left[\frac{\mathrm{pdet}(U_{1,o}\Lambda_{1,o}U_{1,o}^\top + U_{2,o}\Lambda_{2,o}U_{2,o}^\top)}{\sqrt{\det\Lambda_{1,o}\det\Lambda_{2,o}}} \cdot \sqrt{\det\Lambda_{1,n}\cdot\det\Lambda_{2,n}}\right]^{-\frac{1}{2}}$$

where $\mathrm{pdet}(\cdot)$ denotes the pseudo-determinant.

Remark 1. *This misclassification probability has order $\frac{d-r}{2}$. That is, $\lim_{\sigma^2\to 0}\frac{\log\bar{P}_e}{\log\sigma^2} = \frac{d-r}{2}$. The larger the principal angles, the smaller is the upper bound. And misclassification probability decreases most rapidly when the two classes are disjoint ($r = 0$).*

4. STRUCTURE-PRESERVING AND DISCRIMINATIVE FEATURE EXTRACTION

In Theorem 1, it is the product of the sines of the principal angles that determines the performance of the MAP classifier. This motivates an approach to learning features when labels are available from training data. The next section describes how to learn a linear transform that increases separation between classes, allowing for the possibility that the transform reduces dimensionality.

4.1. TRAIT Algorithm

We denote the collection of all labeled training samples as $X = [X_1, \ldots, X_K] \in \mathbb{R}^{n\times N}$, where columns in the submatrix $X_k \in \mathbb{R}^{n\times N_k}$ are samples from the k-th class. The signal subspace of X_k is spanned by the orthonormal basis U_k defined above. The linear transform $A \in \mathbb{R}^{m\times n}$ ($m \le n$) is designed to maximize separation of the subspaces $AU_1, \ldots, AU_K$.

The largest separation is achieved when $(AU_j)^\top(AU_k) = 0$ for all $j \ne k$. In this case, all the principal angles are $\pi/2$. One could use SVD to compute the U_k's first and then learn A. However, in practice we may want to avoid precomputing the U_k's. This can be achieved by encouraging $(AX_j)^\top(AX_k) = 0$ for all $j \ne k$.

We also require the transform A to preserve some specific characteristic or trait of each individual class. For example, we may target $(AX_k)^\top(AX_k) = X_k^\top X_k$ for all k, so that the original intra-class data structure (with noise) is preserved. Given access to denoised signal, $\tilde{X}_k$, we might instead target $(AX_k)^\top(AX_k) = \tilde{X}_k^\top\tilde{X}_k$ again for all k. In this case, the intra-class dispersion due to noise is suppressed. Thus, the Gram matrix T of the transformed signal can be designed to target preservation of particular intra-class structure. We formulate the optimization problem as

$$\min_{A\in\mathbb{R}^{m\times n}} \|(AX)^\top(AX) - T\|_F^2. \tag{4}$$

The block diagonal structure of the target Gram matrix T promotes larger principal angles between subspaces. At the same time the diagonal blocks can be tuned to different characteristics of individual classes. Here we only consider

$$T = \mathrm{diag}\{X_1^\top X_1, \ldots, X_N^\top X_N\}, \tag{5}$$

to demonstrate proof of concept. We refer to this approach as the TRAIT algorithm since it achieves **T**unable **R**ecognition **A**dapted to **I**ntra-class **T**argets.

It is possible to minimize the objective in (4) by first solving $\|X^\top PX - T\| = 0$ for P, and then factoring P as $P = A^\top A$ where $A \in \mathbb{R}^{m\times n}$. However when $m < n$, such a rank-m decomposition may not exist since this P is not guaranteed to be rank deficient. Considering the above fact, here we solve (4) via gradient descent as summarized in Algorithm 1.

4.2. Related Methods

Linear Discriminant Analysis (LDA) is a classical feature extraction method which assumes each class to be Gaussian

Algorithm 1 TRAIT for feature extraction

Input: labeled training samples $X = [X_1, \ldots, X_K]$, target dimension m, $(m \leq n)$, target Gram matrix T.

Output: feature extraction matrix (transform) $A \in \mathbb{R}^{m \times n}$

1: Initialize $A = [e_1, \ldots, e_m]^\top$, where e_i is the i-th standard basis.

2: **while** stopping criteria not met **do**

3: Compute gradient

$$G = A(XX^\top A^\top AXX^\top - XTX^\top).$$

4: Choose a positive step-size η and take a gradient step

$$A \leftarrow A - \eta G.$$

5: **end while**

distributed. It achieves better performance on face recognition tasks than does PCA [14]. LDA does not assume near low-rank structure of the covariances, and therefore considers a different data geometry than the one here studied. Recently, methods of feature extraction based on random projection have been developed and successfully applied to face recognition [15]. The random projection is designed to preserve pairwise distances between all data points uniformly across class labels [16].

More recently, the Low-Rank Transform (LRT) has been proposed as a method of extracting features [3]. It enlarges inter-class distance while suppressing intra-class dispersion. LRT uses the nuclear norm, $\|AX_i\|_*$, to measure the dispersion of the (transformed) data. The transform A is

$$\underset{A \in \mathbb{R}^{m \times n} : \|A\|_2 \leq c}{\arg\min} \sum_{i=1}^{N} \|AX_i\|_* - \|AX\|_*. \tag{6}$$

Note that LRT takes no account of intra-class structure.

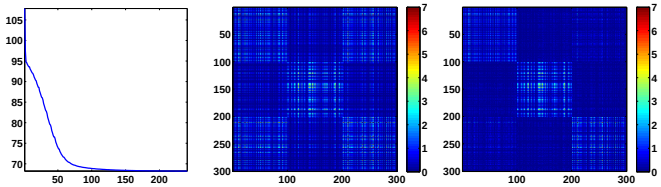
5. EXPERIMENTAL RESULTS

We compare TRAIT with the methods discussed above on both synthetic and measured data. On synthetic data, since the distribution of each class is exactly known, we use a MAP classifier to assess the effectiveness. On measured data, we use the Nearest Subspace Classifier (NSC) instead. The NSC infers the class label of a test datum x as $\arg\min_k \|x - U_k U_k^\top x\|$ where U_k is the orthonormal basis of the k-th signal subspace that can be learned from training data. We do not use a MAP classifier since in most real applications, the distribution of each class is unknown. Even if we assume the distribution to be Gaussian, the number of training samples may be too small to accurately estimate its covariance. On the other hand, NSC only requires the number of training samples to be larger than the assumed subspace dimension. It has been effectively applied in [17, 18] and is reasonable to serve as a benchmark.

5.1. TRAIT on Synthetic Data

The synthetic dataset has parameters $n = 10$, $d = 1$ and $K = 3$. $\Sigma_k = U_k U_k^\top + 10^{-2} U_{k,\perp} U_{k,\perp}^\top (k = 1, 2, 3)$, where U_k is a normalized n-vector with i.i.d. Gaussian random entries. Samples of the k-th class are i.i.d drawn from $\mathcal{N}(0, \Sigma_k)$. For each class, 100 samples are used for learning the transform and 10000 are used for testing. On the training data, we learn the transform respectively via LDA, LRT, and TRAIT with target dimension $m = 3, \ldots, 10$. Then on each test datum, we apply the learned transforms as well as random projection (each entry drawn from $\mathcal{N}(0, 1)$) and classify using a MAP classifier.

We first inspect the iterations of TRAIT as in Algorithm 1. Figure 1a demonstrates the monotonically decreasing objective when $m = 3$. Figures 1b and 1c compare the absolute valued Gram matrices of the training samples before and after TRAIT learns the transform. As expected, the post-transform Gram matrix has an obvious block diagonal structure with the original diagonal blocks roughly preserved. Note that the off-diagonal block values in Figure 1c are significantly suppressed for better inter-class separation.

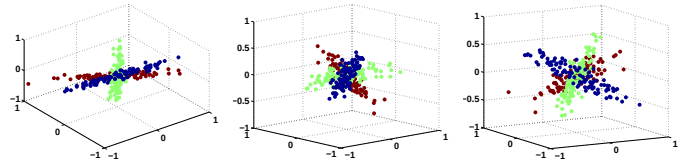


(a) (b) (c)

Fig. 1: TRAIT with target dimension $m = 3$. (a) Objective function value of TRAIT during the iterations of Algorithm 1; (b) Absolute valued Gram matrix of the original training samples; (c) Absolute valued Gram matrix of the transformed training samples by TRAIT.

We then embed the original and transformed data into 3 dimensional coordinates via PCA. Figure 2 demonstrates the embeddings when the target feature dimension is $m = 3$. Each class is represented with a different color. Annotated under each embedding are the angles between the three subspaces of the original/transformed data. TRAIT significantly increases the angles between different subspaces, competitive w.r.t. LRT, while LDA and random projection do not succeed in increasing the separation.

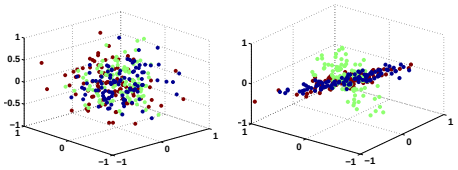
In Figure 3 we vary feature dimension m and compare the MAP classifier's error probabilities on features extracted by different methods. Features extracted by TRAIT and LRT achieve comparable P_e over all the m, outperforming LDA



(a) Original, angles: 66.1°, 21.8°, 70.0°

(b) TRAIT, angles: 87.8°, 72.9°, 88.7°

(c) LRT, angles: 86.0°, 77.1°, 83.0°



(d) LDA, angles: 47.0°, 86.4°, 87.8°

(e) Random, angles: 72.4°, 2.5°, 73.0°

Fig. 2: Embeddings of original and transformed data.

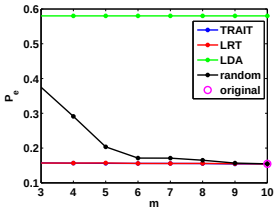


Fig. 3: MAP classifier's P_e on transformed data. Note that TRAIT (blue) almost overlaps with LRT (red).

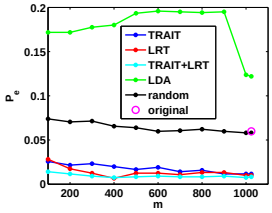


Fig. 4: NSC's P_e on original/transformed face images.

and random projection. Note also that with dimension reduction, TRAIT is able to achieve misclassification probability comparable to that achievable with the initial higher dimensional data.

5.2. TRAIT on Face Dataset

The extended Yale B face database includes 38 subjects, each with 64 images taken under different illumination conditions. We use a cropped version of this data set[1], where each image is of size $32 \times 32 = 1024$. It is known that a 9 dimensional subspace is sufficient to capture the geometry of each subject [1]. In this experiment, we randomly select half of the 64 images from each subject as training and the rest as testing. We vary the target dimension m for all the feature extraction methods and apply NSC to the transformed testing data. The classification error P_e is as shown in figure 4. The NSC achieves much higher accuracies on TRAIT and LRT extracted features than on those of the other methods. Moreover, TRAIT and LRT extracted features are quite different (as will be seen below), suggesting that there is information present in one view that is not present in the other. This is confirmed by noting that when NSC is applied to the concatenation of these two views, the classification accuracy is increased.

The difference between TRAIT and LRT extracted features is due to the intra-class structure preserving property of TRAIT. We illustrate this property by viewing the transformed classes as faces in the original image domain. In figure 5, we display the original images of subject 10 and their and their TRAIT and LRT transforms. TRAIT preserves the various illumination conditions in different images, whereas LRT collapses all images of one subject onto a similar image regardless of their differences. In summary, LRT sacrifices intra-class structure to take maximal advantage of differences in the support of the low-rank subspaces that describe the different classes. LRT and TRAIT are similar in that the block-diagonal target structure in TRAIT penalizes large entries outside the diagonal blocks. However TRAIT is also able to incorporate intra-class structure in the block diagonal target, and classification performance is improved by using LRT features and TRAIT features in combination.

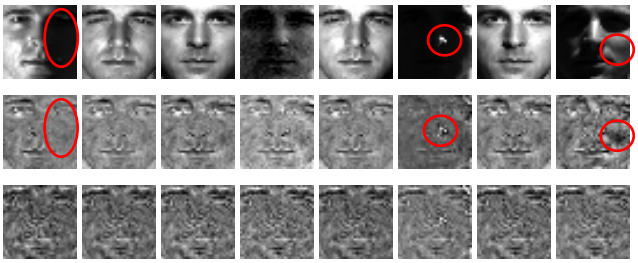


Fig. 5: Comparison of transformed images. Top: original images; Middle: transformed by TRAIT; Bottom: transformed by LRT. Red circles highlight several examples of the preserved structural information from original image to TRAIT transformed ones.

6. CONCLUSION

Starting from a low-rank subspace model, we have shown that the error probability of an optimal classifier is determined by the product of the sines of the principal angles between subspaces. We have designed an algorithm, TRAIT, for extracting low dimensional features, that increases separation between classes, and that can also be tuned to preserve specific intra-class structure. Classification based on TRAIT features alone is competitive with state-of-the-art methods such as LRT. Classification performance is improved by using LRT features and TRAIT features in combination. Future work will include convergence analysis of TRAIT and the design of task-specific target matrices.

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

7. REFERENCES

[1] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[2] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] Q. Qiu and G. Sapiro, "Learning transformations for clustering and classification," *to appear in Journal of Machine Learning Research*, 2014.

[4] L. Applebaum, S. D. Howard, S. Searle, and R. Calderbank, "Chirp sensing codes: Deterministic compressed sensing measurements for fast recovery," *Applied and Computational Harmonic Analysis*, vol. 26, no. 2, pp. 283–290, 2009.

[5] R. Calderbank and S. Jafarpour, "Reed muller sensing matrices and the lasso," *Sequences and Their Applications (SETA), Springer Berlin Heidelberg*, pp. 442–463, 2010.

[6] X. Yang, H. Fu, H. Zha, and J. Barlow, "Semi-supervised nonlinear dimensionality reduction," *Proceedings of the 23rd international conference on Machine learning*, 2006.

[7] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Transaction on Information Theory*, vol. 51, no. 4, pp. 1261–1283, 2005.

[8] A. Ashok, P. K. Baheti, and M. A. Neifeld, "Compressive imaging system design using task-specific information," *Applied Optics*, vol. 47, no. 25, pp. 4457–4471, 2008.

[9] F. Renna, R. Calderbank, L. Carin, and M. Rodrigues, "Reconstruction of signals drawn from a Gaussian mixture via noisy compressive measurements," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2265 – 2277, 2014.

[10] L. Wang, D. Carlson, M. Rodrigues, R. Calderbank, and L. Carin, "A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2611–2629, 2014.

[11] M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications inspired linear discriminant analysis," in *International Conference of Machine Learning*, 2012.

[12] T. Wimalajeewa, H. Chen, and P. K. Varshney, "Performance limits of compressive sensing-based signal classification," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2758–2770, 2012.

[13] H. Reboredo, F. Renna, R. Calderbank, and M. Rodrigues, "Compressive classification of a mixture of gaussians: Analysis, designs and geometrical interpretation," *arXiv preprint arXiv:1401.6962*, 2014.

[14] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[16] W. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," in *Conference in Modern Analysis and Probability*, 1982.

[17] R. Basri, T. Hassner, and L. Zelnik-Manor, "Approximate nearest subspace search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 266–278, 2011.

[18] Y. Liu, S. S. Ge, C. Li, and Z. You, "K-ns: A classifier by the distance to the nearest subspace," *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1256–1268, 2011.