LINEAR DISCRIMINANT ANALYSIS WITH FEW TRAINING DATA

Panos P. Markopoulos

Department of Electrical and Mircoelectronic Engineering
Rochester Institute of Technology
Rochester, NY 14623
E-mail: panos@rit.edu

ABSTRACT

Statistically-optimal Linear Discriminant Analysis (LDA) is formulated as a maximization that involves the nominal statistics of the classes to be discriminated. In practice, however, these nominal statistics are unknown and estimated from a collection of labeled training data. Accordingly, the nominal LDA basis is approximated by the solution of the popular practical LDA problem defined upon these estimates. However, when the available training data are few, the solution to practical LDA is known to lie far from the nominal LDA basis. In this work, we propose a novel algorithm that operates on the estimated class statistics and generates a sequence of bases that converges to the solution of practical LDA. Importantly, our studies illustrate that early elements of the proposed sequence exhibit significantly higher approximation to the nominal LDA basis than the converging point and, thus, offer the means for superior classification performance.

Index Terms— Dimensionality reduction, classification, linear discriminant analysis, recognition, subspace learning.

1. INTRODUCTION AND PROBLEM STATEMENT

Linear Discriminant Analysis (LDA) is the fundamental data analysis method, introduced in [1], that has been used extensively in the past decades for dimensionality reduction, recognition, and supervised classification. LDA finds application in a wide range of fields such as Image Processing, Computer Vision, Pattern Recognition, and Bioinformatics [2–11], to name a few. Some of the advantages of LDA that have contributed to its immense popularity are (i) its low-cost implementation (solution through simple generalized eigenvalue decomposition), (ii) its correspondence to Bayes's optimal classification, for two homoscedastic Gaussian classes with equal priors, and (iii) its easy adaptation for discriminating non-linearly separable classes, through the kernel trick method [12, 13].

The notion behind LDA is to identify a low-dimensional linear subspace whereon the data points of two or more classes are best separable. Mathematically, given C classes of D-dimensional points, multi-class LDA seeks for a basis $\mathbf{W}^* = [\mathbf{w}_1^*, \mathbf{w}_2, \ldots, \mathbf{w}_K^*] \in \mathbb{R}^{D\times K}$ that describes a K-dimensional subspace whereon projected data exhibit high between-class separation and low within-class dispersion. That is, if elements in class c are distributed by $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, for $c = 1, 2, \ldots, C$, multi-class LDA seeks the rank-K solution $\mathbf{W}^* \in \mathbb{R}^{D\times K}$ to the determinant-quotient problem [14, 15]

$$\underset{\mathbf{W}\in\mathbb{R}^{D\times K};\ \mathbf{W}^\top\mathbf{W}=\mathbf{I}_K}{\max.} J_K(\mathbf{W}) := \frac{|\mathbf{W}^\top\boldsymbol{\Sigma}_b\mathbf{W}|}{|\mathbf{W}^\top\boldsymbol{\Sigma}\mathbf{W}|} \quad (1)$$

where $|\cdot|$ denotes the determinant of the matrix argument, $\boldsymbol{\Sigma}_b$ is the between-class covariance matrix defined as $\boldsymbol{\Sigma}_b := \frac{1}{C}\sum_{c=1}^{C}(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top$, and $\boldsymbol{\mu}$ is the mean of the class means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_C$. A solution to (1) is known to be given by the K eigenvectors of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_b$ that correspond to the K highest, non-zero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_K$ [15]; that is, by $\mathbf{W}^*$ such that $\|\mathbf{w}_k^*\|_2 = 1$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_b\mathbf{w}_k^* = \lambda_k\mathbf{w}_k^*\ \forall k$.

In practice, however, the nominal class statistics $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, …, $\boldsymbol{\mu}_C$, and $\boldsymbol{\Sigma}$ are unknown to the classifier and *ideal LDA* in (1) cannot be formulated. Instead, a common practical approach is to estimate these statistics from a collection of labeled training data from each class. That is, given N_c training points $\{\mathbf{x}_n^{(c)}\}_{n=1}^{N_c}$ from class c, $c = 1, 2, \ldots, C$, $\boldsymbol{\mu}_c$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_b$ are estimated by

$$\hat{\boldsymbol{\mu}}_c := \frac{1}{N_c}\mathbf{X}_c\mathbf{1}_{N_c},\ \ \hat{\boldsymbol{\mu}} := \frac{1}{N}\sum_{c=1}^{C} N_c\hat{\boldsymbol{\mu}}_c, \quad (2)$$

$$\hat{\boldsymbol{\Sigma}} := \frac{1}{N}\sum_{c=1}^{C}\sum_{n=1}^{N_c}(\mathbf{x}_n^{(c)} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_n^{(c)} - \hat{\boldsymbol{\mu}}_c)^\top, \text{ and} \quad (3)$$

$$\hat{\boldsymbol{\Sigma}}_b := \frac{1}{N}\sum_{c=1}^{C} N_c(\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}})^\top, \quad (4)$$

respectively, where $N := N_1 + N_2 + \ldots + N_C$. Accordingly, the solution to (1) is approximated by the solution $\hat{\mathbf{W}} \in \mathbb{R}^{D\times K}$ to the *practical LDA* problem

$$\underset{\mathbf{W}\in\mathbb{R}^{D\times K};\ \mathbf{W}^\top\mathbf{W}=\mathbf{I}_K}{\max.} \hat{J}_K(\mathbf{W}) := \frac{|\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}_b\mathbf{W}|}{|\mathbf{W}^\top\hat{\boldsymbol{\Sigma}}\mathbf{W}|}, \quad (5)$$

given by the K highest-eigenvalue eigenvectors of $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_b$.

Understandably, the quality of $\hat{\mathbf{W}}$ with respect to ideal LDA metric in (1) and, therefore, its nominal class discriminating capability, depend on the quality of the estimates in (2)-(4). When the number of training points $\{N_c\}_{c=1,2,\ldots,C}$ available from each class is asymptotically large ($N \gg D$), estimates in (2)-(4) tend to the nominal class statistics, $\hat{J}(\mathbf{W})$ tends to $J(\mathbf{W})$, for every $\mathbf{W} \in \mathbb{R}^{D\times K}$, and the solution to (5), $\hat{\mathbf{W}}$, approximates the ideal LDA basis $\mathbf{W}^*$.

In many cases of practical interest, however, where D is large, the training points available to the classifier are not enough for estimating accurately the class statistics and, in particular, the within-class covariance matrix $\mathbf{\Sigma}$ (*short-training-data* case). Examples include many image processing applications, such as object/face recognition/classification [16, 17], where $h \times w$ images are treated as $(D = hw)$-dimensional points (e.g., for $h = w = 128$, $D = 16\,384$). In such cases the collection of $N \gg D$ labeled training points becomes infeasible (or, at least, impractical) and, even if $N > D$, $\hat{\mathbf{\Sigma}}$ is not an accurate estimate of $\mathbf{\Sigma}$. Accordingly, $\hat{J}(\mathbf{W})$ diverges from $J(\mathbf{W})$ and its maximizer, $\hat{\mathbf{W}}$, may lie significantly far from the sought-after basis $\mathbf{W}^*$. In the extreme case where $N < D$, the well-known LDA (covariance-matrix) singularity problem emerges and, as a remedy, a small, heuristically chosen regularizer $\Delta > 0$ is commonly added to the diagonal elements of $\hat{\mathbf{\Sigma}}$ [3, 18]; this approach resolves the singularity issue, but does not necessarily render $\hat{\mathbf{\Sigma}}$ an accurate estimate for $\mathbf{\Sigma}$.

In this work, inspired by relevant successful developments in the auxiliary-vector (AV) filtering literature [19–22], we propose a novel iterative algorithm for the estimation of the ideal LDA basis $\mathbf{W}^*$ from short training data. The proposed algorithm operates directly on the estimates $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}_b$ in (3)-(4) and generates a sequence of LDA bases $\{\mathbf{W}_t\}_{t=1,2,\ldots}$ that provably converges to the solution of (5), $\hat{\mathbf{W}}$. What is most interesting, however, is that early elements of the generated basis sequence are shown to exhibit significantly improved approximation to the ideal basis $\mathbf{W}^*$, compared to $\hat{\mathbf{W}}$, in the short-training-data case; thus, they offer the grounds for LDA-classification of superior performance.

2. PROPOSED ITERATIVE LINEAR DISCRIMINANT ANALYSIS

We commence our developments by defining $\mathbf{V} := [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_C] \in \mathbb{R}^{D\times C}$, where $\mathbf{v}_c := \sqrt{\frac{N_c}{N}}(\hat{\boldsymbol{\mu}}_c - \hat{\boldsymbol{\mu}})$ for every c, such that $\hat{\mathbf{\Sigma}}_b = \mathbf{V}\mathbf{V}^\top$. Then, we denote by $\mathbf{Z} := [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K] \in \mathbb{R}^{C\times K}$ the matrix containing the K eigenvectors of $\mathbf{V}^\top \hat{\mathbf{\Sigma}}^{-1}\mathbf{V}$, corresponding to its K highest eigenvalues $p_1, p_2, \ldots, p_K$. Accordingly, we define the $K \times K$ diagonal matrix $\mathbf{P} := \mathrm{diag}([p_1, p_2, \ldots, p_K]^\top)$. In view of the above definitions, it can be easily shown that for the solution to (5), $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots \hat{\mathbf{w}}_K] \in \mathbb{R}^{D\times K}$, given by the K eigenvectors of $\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_b$ that correspond to its highest non-zero eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$, it holds

$$\hat{\mathbf{w}}_k = \frac{\hat{\mathbf{\Sigma}}^{-1}\mathbf{V}\mathbf{z}_k}{\|\hat{\mathbf{\Sigma}}^{-1}\mathbf{V}\mathbf{z}_k\|_2} \;\; \forall k. \tag{6}$$

This alternative formulation of the solution to (5) lies in the core of our algorithmic developments presented below.

Proposed Iterative LDA Algorithm

Initialization:
$\mathbf{b}_{0,k} \leftarrow \mathbf{V}\mathbf{z}_k \;\; \forall k, \;\; \mathbf{w}_{0,k} \leftarrow \frac{\mathbf{b}_{0,k}}{\|\mathbf{b}_{0,k}\|_2} \;\; \forall k$
$\mathbf{P}_k \leftarrow \mathbf{I}_D - \|\mathbf{V}\mathbf{z}_k\|_2^{-2}\mathbf{V}\mathbf{z}_k\mathbf{z}_k^\top\mathbf{V}^\top \;\; \forall k$
Iterations:
for $t = 1, 2, \cdots$
  for $k = 1, 2, \cdots, K$
    $\mathbf{p} \leftarrow \mathbf{P}_k\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}$;
    if $\|\mathbf{p}\|_2 > 0$
      $\mathbf{g}_{t,k} \leftarrow \frac{\mathbf{p}}{\|\mathbf{p}\|_2}, \;\; \omega_{t,k} \leftarrow \frac{\mathbf{g}_{t,k}^\top\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}}{\mathbf{g}_{t,k}^\top\hat{\mathbf{\Sigma}}\mathbf{g}_{t,k}}$
      $\mathbf{b}_{t,k} \leftarrow \mathbf{b}_{t-1,k} - \omega_{t,k}\mathbf{g}_{t,k}$
    else
      $\mathbf{b}_{t,k} \leftarrow \mathbf{b}_{t-1,k}$
    $\mathbf{w}_{t,k} \leftarrow \frac{\mathbf{b}_{t,k}}{\|\mathbf{b}_{t,k}\|_2}$
  If convergence is met for all k, *break*;
Return: LDA-basis sequence $\{\mathbf{W}_t\}_{t=1,2,\ldots}$.

Fig. 1: The proposed algorithm that generates LDA-basis sequence $\{\mathbf{W}_t\}_{t=1,2,\ldots}$.

The proposed LDA algorithm runs iteratively. At each iteration step, say the tth, it generates a subspace basis $\mathbf{B}_t = [\mathbf{b}_{t,1}, \mathbf{b}_{t,2}, \ldots, \mathbf{b}_{t,K}] \in \mathbb{R}^{D\times K}$ and normalizes its columns, in accordance to (6), to deliver an approximate LDA basis $\mathbf{W}_t$; that is, for every iteration step $t \geq 0$,

$$\mathbf{w}_{t,k} := \frac{\mathbf{b}_{t,k}}{\|\mathbf{b}_{t,k}\|_2} \;\; \forall k. \tag{7}$$

In view of (7), in the sequel we focus on describing the iterative generation of sequence $\{\mathbf{B}_t\}_t$. For $t = 0$, we initialize the iterations at $\mathbf{B}_0$, with

$$\mathbf{b}_{0,k} := \frac{\mathbf{V}\mathbf{z}_k}{\|\mathbf{V}\mathbf{z}_k\|_2^2} \;\; \forall k, \tag{8}$$

motivated by the fact that for $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}$, $\hat{\mathbf{\Sigma}}_b = \mathbf{\Sigma}_b$, and $\mathbf{\Sigma} = \mathbf{I}_D$, the kth initial basis vector $\mathbf{w}_{0,k} = \frac{\mathbf{b}_{0,k}}{\|\mathbf{b}_{0,k}\|_2}$ coincides with the ideal $\mathbf{w}_k^*$, for every k. Thereafter, at the general iteration step $t \geq 1$, $\mathbf{b}_{t,k}$ is generated by incorporating to $\mathbf{b}_{t-1,k}$ a vector component perpendicular to the initial direction $\mathbf{V}\mathbf{z}_k$, as

$$\mathbf{b}_{t,k} := \mathbf{b}_{t-1,k} - \omega_{t,k}\mathbf{g}_{t,k} \;\; \forall k, \tag{9}$$

where $\mathbf{g}_{t,k} \in \mathbb{R}^{D\times 1}$, $\|\mathbf{g}_{t,k}\|_2 = 1$, $\mathbf{g}_{t,k}^\top\mathbf{V}\mathbf{z}_k = 0$, and $\omega_{t,k} \in \mathbb{R}$. Considering for a moment direction vector $\mathbf{g}_{t,k}$ to be fixed and non-zero, we design parametrically the scaling coefficient

$\omega_{t,k}$ so that the within-class dispersion of training data components on $\mathbf{b}_{t,k}$ is minimized; that is, $\omega_{t,k}$ is defined as the solution to

$$\min_{\omega\in\mathbb{R}}.\ (\mathbf{b}_{t-1,k}+\omega\mathbf{g}_{t,k})^\top\hat{\mathbf{\Sigma}}(\mathbf{b}_{t-1,k}+\omega\mathbf{g}_{t,k}). \tag{10}$$

Setting to zero the derivative of (10) with respect to ω, scaling coefficient $\omega_{t,k}$ is found to be

$$\omega_{t,k} := \frac{\mathbf{g}_{t,k}^\top\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}}{\mathbf{g}_{t,k}^\top\hat{\mathbf{\Sigma}}\mathbf{g}_{t,k}}\ \ \forall k. \tag{11}$$

Then, we steer our focus to the direction-vector $\mathbf{g}_{t,k}$ and define it as the normalized vector in $\mathbb{R}^{D\times 1}$, orthogonal to $\mathbf{V}\mathbf{z}_k$, that lies closest to $\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}$. That is, we define $\mathbf{g}_{t,k}$ as the solution to

$$\min_{\mathbf{g}\in\mathbb{R}^{D\times 1};\ \mathbf{g}^\top\mathbf{V}\mathbf{z}_k=0;\ \|\mathbf{g}\|_2=1}.\ \|\mathbf{g}-\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}\|_2^2. \tag{12}$$

Solving (12) by Lagrange multipliers, we find that, if $\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}$ is not a scaled version of the initialization vector $\mathbf{V}\mathbf{z}_k$, then

$$\mathbf{g}_{t,k} := \frac{\left(\mathbf{I}_D-\frac{\mathbf{V}\mathbf{z}_k\mathbf{z}_k^\top\mathbf{V}^\top}{\|\mathbf{V}\mathbf{z}_k\|_2^2}\right)\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}}{\|\left(\mathbf{I}_D-\frac{\mathbf{V}\mathbf{z}_k\mathbf{z}_k^\top\mathbf{V}^\top}{\|\mathbf{V}\mathbf{z}_k\|_2^2}\right)\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}\|_2}\ \ \forall k. \tag{13}$$

If, otherwise, there exists $\alpha\in\mathbb{R}$ such that $\hat{\mathbf{\Sigma}}\mathbf{b}_{t-1,k}=\alpha\mathbf{V}\mathbf{z}_k$, then $\mathbf{g}_{t,k}$ can take any value of unit magnitude that is orthogonal to $\mathbf{V}\mathbf{z}_k$, yielding a maximum value of zero in (12); evidently, in this case, $\omega_{t,k}$ in (11) becomes zero, $\mathbf{b}_{t,k}$ coincides with $\mathbf{b}_{t-1,k}$, and the sequence the kth column converges.

The following Lemmata 1 and 2 describe formally the convergence of the proposed LDA-basis sequence – the corresponding proofs are omitted from this manuscript due to lack of space.

Lemma 1 *For every $k\in\{1,2,\ldots,K\}$, it holds that*

$$\lim_{t\to\infty}\left(\mathbf{I}_D-\frac{\mathbf{V}\mathbf{z}_k\mathbf{z}_k^\top\mathbf{V}^\top}{\|\mathbf{V}\mathbf{z}_k\|_2^2}\right)\hat{\mathbf{\Sigma}}\mathbf{b}_{t,k}=\mathbf{0}_D. \tag{14}$$

By Lemma 1, (11) and (12), $\omega_{t,k}$ converges to zero, for every column-index k, as t tends to infinity. Therefore, the iteratively generated sequence $\{\mathbf{b}_{t,k}\}_{t=1,2,\ldots}$ in (9) converges, for every k; this in turn implies that both $\{\mathbf{B}_t\}_{t=1,2,\ldots}$ and the proposed LDA-basis sequence $\{\mathbf{W}_t\}_{t=1,2,\ldots}$ converge as well. The exact converging point of $\{\mathbf{b}_{t,k}\}_{t=1,2,\ldots}$ is presented in the following Lemma 2.

Lemma 2 *For every $k\in\{1,2,\ldots,K\}$, it holds that*

$$\lim_{t\to\infty}\mathbf{b}_{t,k}=\frac{1}{p_k}\hat{\mathbf{\Sigma}}^{-1}\mathbf{V}\mathbf{z}_k\ \ \forall k. \tag{15}$$

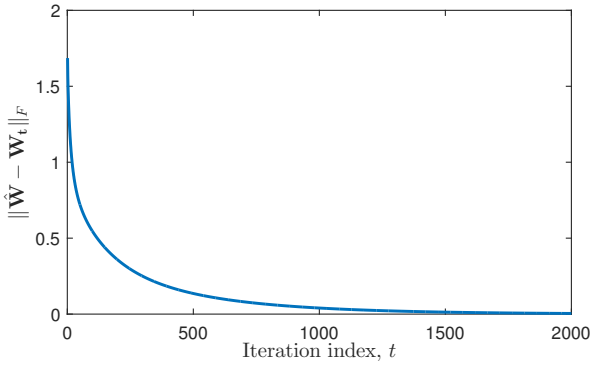

Fig. 2: Convergence of the proposed LDA-basis sequence $\{\mathbf{W}_t\}_{t=1,2,\ldots}$ to $\hat{\mathbf{W}}$, solution to (5), captured by $\|\hat{\mathbf{W}}-\mathbf{W}_t\|_F$ ($D=40$, $C=3$, $K=2$, $N_1=N_2=N_3=20$).

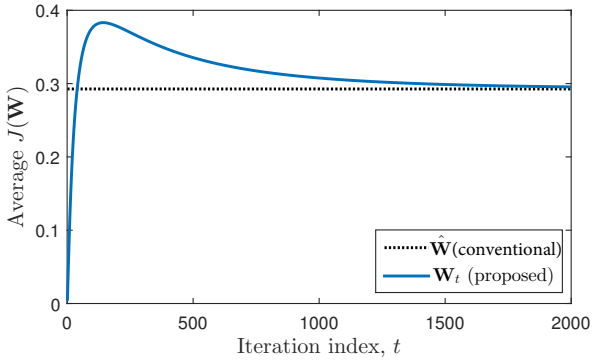

Fig. 3: Average performance of the proposed LDA-basis sequence $\{\mathbf{W}_t\}_{t=1,2,\ldots}$, with respect to the LDA metric in (1), along benchmark performance of $\hat{\mathbf{W}}$, solution to (5) ($D=40$, $C=3$, $K=2$, $N_1=N_2=N_3=20$).

We notice now that the converging point of $\{\mathbf{b}_{t,k}\}_{t=1,2,\ldots}$ is a scaled version of $\hat{\mathbf{w}}_k$ in (6). Therefore, in view of the normalization step in (7), the convergence of the proposed LDA-basis sequence formally converges to the solution of (5); i.e.,

$$\lim_{t\to\infty}\mathbf{W}_t=\hat{\mathbf{W}}. \tag{16}$$

The convergence of $\{\mathbf{W}_t\}_{t=1,2,\ldots}$ to $\hat{\mathbf{W}}$ is illustrated in Fig. 2, by means of $\|\hat{\mathbf{W}}-\mathbf{W}_t\|_F$, $t=1,2,\ldots,2000$, for a problem instance of $D=40, C=3, K=2$, and $N_1=N_2=N_3=20$. For completeness, a detailed pseudo-code description of the proposed algorithm is offered in Fig. 1.

3. EXPERIMENTAL STUDIES AND CONCLUSIONS

As a first evaluation study of the performance of the proposed LDA-basis sequence, we fix arbitrarily class statistics $\mathbf{\Sigma}$, and $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_C$, for $D=40$, $C=3$, and $K=2$, and generate 1000 independent training sets of $N=60$ points each ($N_1=N_2=N_3=20$). For each training set, we estimate independently the class statistics by (2)-(4). Then, we calculate the solution to (5), $\hat{\mathbf{W}}$, generate the 2000 first elements of

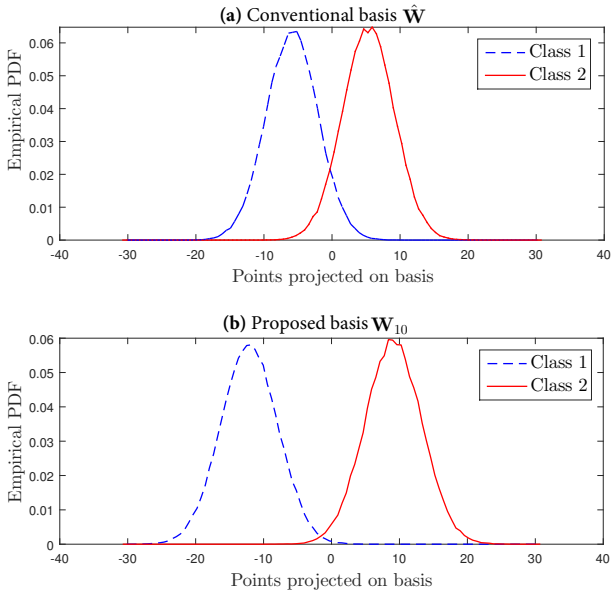

Fig. 4: PDF of data projections on (a) the conventional basis $\hat{\mathbf{W}}$, solution to (5), and (b) the 10th element of the proposed sequence $\mathbf{W}_{10}$ ($D = 50, C = 2, K = 1, N_1 = N_2 = 30$).

the proposed basis sequence, $\{\mathbf{W}_t\}_{t=1,2,\ldots,2000}$, and evaluate their corresponding performances in the metric of (1) ($J(\hat{\mathbf{W}})$ and $\{J(\mathbf{W}_t)\}_{t=1,2,\ldots,2000}$). In Fig. 3, we plot the average values of $J(\hat{\mathbf{W}})$ and $\{J(\mathbf{W}_t)\}_{t=1,2,\ldots,2000}$ calculated over the 1000 independent training sets. We observe that, in accordance to our theoretical instantaneous convergence studies above, as t tends to 2000 the performance of the proposed sequence converges to that of $\hat{\mathbf{W}}$. However, quite interestingly, $\{J(\mathbf{W}_t)\}_{t=1,2,\ldots,2000}$ converges to $J(\hat{\mathbf{W}})$ *from above* so that all elements of $\{\mathbf{W}_t\}_{t=1,2,\ldots}$ from $t = 41$ to $t = 2000$ exhibit average performance superior to that of $\hat{\mathbf{W}}$.

Next, we illustrate the class-discrimination merit of the proposed sequence. We consider $C = 2$ classes of $(D = 50)$-dimensional points and seek the $(K = 1)$-dimensional subspace on which classes are best discriminated. The classifier has access to $N_1 = N_2 = 30$ labeled training points from each class. In Fig. 4, we plot the probability-density function (PDF) –empirically calculated over 10 000 points from each class– of the points projected on the conventional basis $\hat{\mathbf{W}} \in \mathbb{R}^{D\times 1}$, solution to (5), and the $(t = 10)$th element of the proposed basis sequence, $\mathbf{W}_{10} \in \mathbb{R}^{D\times 1}$. Interestingly, while $\hat{\mathbf{W}}$ exhibits low class-separation ability (the two PDFs overlap extensively), the 10th element of the proposed sequence, calculated over the same training data as $\hat{\mathbf{W}}$, manages to attain high class separation.

Finally, we conduct a third study on the MNIST Handwritten Digits of [23] that offers training and testing data for $C = 10$ classes of $(D = 784)$-dimensional (vectorized) handwritten digits '0','1', ...,'9'. We use $N_c = 150$ points from each class to train K-dimensional bases $\hat{\mathbf{W}}$ and $\mathbf{W}_{10}$, after regularizing appropriately the singular covariance matrix with

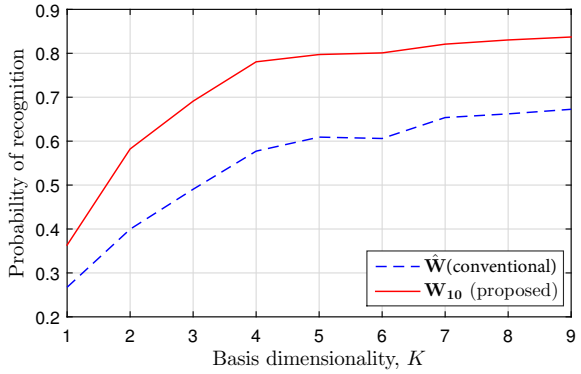

Fig. 5: Probability of digit recognition by means of $\hat{\mathbf{W}}$ and the proposed $\mathbf{W}_{10}$ vs. basis dimensionality $K = 1, 2, \ldots, 9$ ($D = 784, C = 10, N_c = 150$).

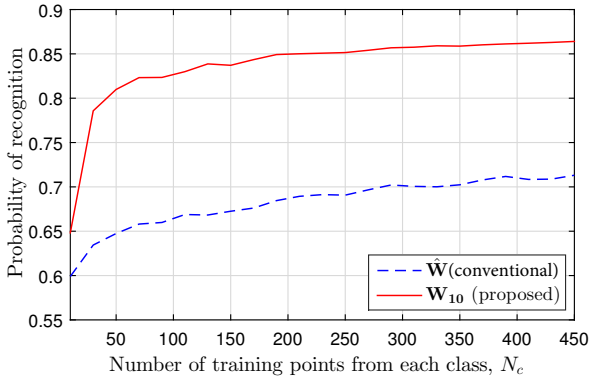

Fig. 6: Probability of digit recognition by means of $\hat{\mathbf{W}}$, solution to (5), and $\mathbf{W}_{10}$ vs. number of training points per class $N_c = 10, 30, \ldots, 450$ ($D = 784, C = 10, K = 150$).

$\Delta = 1$. For each class, we calculate the centroid (mean) of the training data projected on each basis. Then, we project 892 testing points from each class on the calculated bases and apply standard nearest-class-centroid classification. In Fig. 5, we plot the probability of correct digit recognition for $\hat{\mathbf{W}}$ and $\mathbf{W}_{10}$, versus the basis dimensionality $K = 1, 2, \ldots, 9$. Interestingly, for all values of K, the proposed basis $\mathbf{W}_{10}$ outperforms significantly its conventional counterpart, achieving recognition probability of up to 84%, for $K = 9$, while the performance of $\hat{\mathbf{W}}$ does not exceed 75%. In Fig. 6, we keep K fixed to 9 and plot the probability of correct classification for the two bases, versus the number of training samples from each class, $N_c = 10, 30, \ldots, 450$. Once again, the superior classification performance attained by means of proposed basis sequence is clear, achieving recognition probability of 86.5%, for $N_c = 450$.

The presented experimental studies motivate our optimism that the proposed LDA-basis sequence could break new grounds for preferred classification, recognition, and learning tools from few training data.

4. REFERENCES

[1] R.A. Fisher, "The statistical utilization of multiple measurements," *Ann. Eugenics*, vol. 8, pp. 376-386, 1938

[2] E. Alpaydin. *Introduction to machine learning.* Cambridge, MA: MIT press, 2014.

[3] Y. Guo, T. Hastie, R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics,* vol. 8, pp. 86-100, 2007.

[4] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition." *IEEE Trans. Image Process.*, vol. 23, pp. 5599-5611, Dec. 2014.

[5] M.H. Siddiqi, R. Ali, A. M. Khan, P. Young-Tack, and L. Sungyoung, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Trans. Image Process.*, vol. 24, pp. 1386-1398, Apr. 2015.

[6] K. Papachristou, A. Tefas, and I. Pitas, "Facial image analysis based on two-dimensional linear discriminant analysis exploiting symmetry," in *Proc. IEEE Int. Conf. Image Process. (IEEE ICIP 2015)*, Quebec City, Canada, Sep. 2015, pp. 3185-3189.

[7] L. Zhen, L. Shengcai, M. Pietikainen, and S. Z. Li "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.*, vol. 20, pp. 247-256, Jan. 2011.

[8] S. M. Lajevardi and R. W. Hong, "Facial expression recognition in perceptual color space," *IEEE Trans. Image Process.*, vol. 21, pp. 3721-3733, Aug. 2012.

[9] C. Zhen, S. Shiguang, Z. Haihong, L. Shihong, and C. Xilin, "Structured sparse linear discriminant analysis," in *Proc. IEEE Int. Conf. Image Process. (IEEE ICIP 2012)*, Orlando, FL, Sep. 2012, pp. 1161-1164.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Merging linear discriminant analysis with Bag of Words model for human action recognition," in *Proc. IEEE Int. Conf. Image Process. (IEEE ICIP 2015)*, Quebec City, Canada, Sep. 2015, pp. 832-836.

[11] X. Shu, Y. Gao, H. Lu, "Efficient linear discriminant analysis with locality preserving for face recognition," *Patt. Recogn.,* vol. 45, pp. 1892-1898, May 2012.

[12] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. Muller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Neural Net. Signal Process. Workshop,* Madison, WI, Aug. 1999, pp. 41-48.

[13] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Comput.*, vol. 12, pp. 2385-2404, Oct. 2000.

[14] R. C. Rao,"The utilization of multiple measurements in problems of biological classification." *J. Royal Stat. Soc.,* pp. 159-203, Jan. 1948.

[15] K. Fukunaga. *Introduction to statistical pattern recognition*. Cambridge, MA: Academic press, 2013.

[16] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfield, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, pp. 399-458, Dec. 2003.

[17] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-Based algorithms," *IEEE Trans. Neural Net.*, vol. 14, pp. 195-200, Jan. 2003.

[18] A. R. Webb. *Statistical Pattern Recognition.* Hoboken, NJ: John Wiley and Sons, 2002.

[19] D. A. Pados and S. N. Batalama, "Joint space-time auxiliary-vector filtering for DS/CDMA systems with antenna arrays," *IEEE Trans. Commun.*, vol. 47, pp. 1406-1415, Sep. 1999.

[20] D. A. Pados and G. N. Karystinos, "An iterative algorithm for the computation of the MVDR filter," *IEEE Trans. Signal Process.*, vol. 49, pp. 290-300, Feb. 2001.

[21] P. P. Markopoulos, S. Kundu, and D. A. Pados, "Small-sample-support suppression of interference to PN-masked data," *IEEE Trans. Commun.*, vol. 61, pp. 2979-2987, Jul. 2013.

[22] P. P. Markopoulos, S. Kundu, and D. A. Pados, "Short-data-record filtering of PN-masked data," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (IEEE ICASSP 2013)*, May 2013, pp. 4559-4563.

[23] Binary Alphadigits Database. (2006, Oct. 15) [Online]. Available: http://www.cs.nyu.edu/ roweis/data.html.