Identity and Variation Spaces: Revisiting the Fisher Linear Discriminant

Sheng Zhang[†] Terence Sim[‡] Mei-Chen Yeh[§]

[†] Department of Psychology, University of California, Santa Barbara, USA.
[‡] School of Computing, National University of Singapore, Singapore.
[§] Department of Computer Science and Information Engineering, National Taiwan Normal University Taipei, Taiwan.

Abstract

The Fisher Linear Discriminant (FLD) is commonly used in classification to find a subspace that maximally separates class patterns according to the Fisher Criterion. It was previously proven that a pre-whitening step can be used to truly optimize the Fisher Criterion. In this paper, we study the theoretical properties of the subspaces induced by this whitened FLD. Of the four subspaces induced, two are most important for classification and representation of patterns. We call these Identity Space *and* Variation Space. *We show that only the between-class variation remains in Identity Space, and only the within-class variation remains in Variation Space. Both spaces can be used for decomposition and representation of class data. Moreover, we give sufficient conditions for these spaces to exist. Finally, we also run experiments to show how Identity and Variation Spaces may be used for classification and image synthesis.*

1. Introduction

Among linear discriminants, the Fisher Linear Discriminant (FLD) [5], also known as Linear Discriminant Analysis (LDA), is quite possibly the most popular for pattern classification. It has been widely applied in face recognition [3]. FLD is attractive because it is conceptually straightforward and computationally efficient. At the heart is the maximization of the Fisher Criterion:

$$J_F(\mathbf{\Phi}) = trace\{(\mathbf{\Phi}^\top \mathbf{S}_w \mathbf{\Phi})^{-1}(\mathbf{\Phi}^\top \mathbf{S}_b \mathbf{\Phi})\}, \qquad (1)$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between-class and within-class scatter matrices, respectively, and $\mathbf{\Phi}$ is the linear transformation matrix to be found by maximizing J_F. (See Section 2 for the definitions of $\mathbf{S}_b$ and $\mathbf{S}_w$.)

This Criterion expresses the idea that a good feature (computed by projecting the original feature vector onto the subspace of $\mathbf{\Phi}$) should be completely sensitive to variations between classes, while at the same time be completely insensitive to variations within each class. Ideally, all patterns from one class should project onto the same point (resulting in zero within-class variation), while patterns belonging to different classes are projected far away from one another (large between-class variation).

However, as we observed in our previous work [13, 14], most researchers apply the FLD in a sub-optimal manner. That is, the Fisher Criterion J_F is not truly maximized. This often occurs when solving the "singularity problem" incorrectly. The singularity problem refers to the situation when $\mathbf{S}_w$ is singular (non-invertible), and usually occurs when the number of training samples N is less than the dimension D of the feature vectors (called the "small sample-size problem"). In attempting to overcome this singularity problem, many techniques [3, 12] inadvertently discard important discriminative information, rendering the FLD sub-optimal.

Apart from the small sample-size problem, the FLD in general leaves a number of questions unanswered: (a) Does maximizing J_F guarantee perfect class separation? (b) If not, under what conditions can classes be perfectly separated? (c) Can these conditions be satisfied in practice? (d) How are the class patterns distributed in the subspace $\mathbf{\Phi}$? We will answer these questions in this paper.

Our previous paper [13] showed that by first applying the Fukunaga-Koontz Transform (FKT) [7] (essentially a pre-whitening step) to the FLD, the Fisher Criterion J_F can achieve a theoretical maximum value of $+\infty$. This is clearly the best possible value. Moreover, this combination of FKT and FLD decomposed the whole data space into four subspaces, thereby providing additional insight into where common computational errors are made that yield a sub-optimal J_F. However, we did not study the properties of these subspaces, which we now take up in this paper, and which help to answer the above questions.

More precisely, of the four subspaces, two are significant for pattern classification and representation: *Identity*

Space and *Variation Space* (see Sections 3 and 4 for definitions). Our paper explores the theoretical properties of these two subspaces, thereby giving further insight into FLD. We make the following contributions:

1. We prove mathematically that in Identity Space, all classes are perfectly separated: all within-class variation has been "projected out", and only the between-class variation remains. This is what makes the Fisher Criterion achieve its best value of $+\infty$.

2. We further show the geometric structure of Identity Space: the class means are maximally spread out over a hypersphere. They are equidistant from the global mean (*i.e*. the mean of all data, regardless of class). In other words, the class means form the vertices of a regular simplex.

3. We prove analogous properties in Variation Space: (a) all between-class variation has been projected out, and only within-class variation remains; (b) the within-class variation of each class occupies orthogonal subspaces; and (c) the samples in each class are arranged in a regular simplex.

4. We show that Variation Space contains discrimination information, even though all class means are equal. We show how both spaces can be used to decompose and represent class patterns.

2. Mathematical Background

We begin by letting $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, with $\mathbf{x}_i \in \mathbb{R}^D$, denote a dataset of D-dimensional feature vectors. Each feature vector $\mathbf{x}_i$ belongs to exactly one of C classes $\{L_1, ..., L_C\}$. Let $\mathbf{m}_k$ denote the mean of class L_k, and suppose each class has the same number of vectors n, so that $N = nC$. Without loss of generality, we will assume that the global mean of $\mathbf{X}$ is zero, *i.e*. $(\sum_i \mathbf{x}_i)/N = \mathbf{m} = \mathbf{0}$. If not, we may simply subtract $\mathbf{m}$ from each $\mathbf{x}_i$. Define the between-class scatter matrix $\mathbf{S}_b$, the within-class scatter matrix $\mathbf{S}_w$, and the total scatter matrix $\mathbf{S}_t$ as follows:

$$\mathbf{S}_t = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top, \quad \mathbf{S}_b = \sum_{k=1}^{C} N_k \mathbf{m}_k \mathbf{m}_k^\top \tag{2}$$

$$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{\mathbf{x}_i \in L_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top \tag{3}$$

Let $r_t = rank(\mathbf{S}_t)$, $r_b = rank(\mathbf{S}_b)$, and $r_w = rank(\mathbf{S}_w)$. We remark without proof that $\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w$.

2.1. Whitened Fisher Linear Discriminant (WFLD)

To whiten the data, first compute the total scatter matrix $\mathbf{S}_t = \mathbf{X}\mathbf{X}^\top$, then eigen-decompose it to get $\mathbf{S}_t = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, retaining only non-zero eigenvalues in the diagonal matrix $\mathbf{D}$ and their corresponding eigenvectors in $\mathbf{U}$. Now compute the $(N-1) \times D$ matrix $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$, and apply it to the data to get the $(N-1) \times N$ matrix: $\widetilde{\mathbf{X}} = \mathbf{P}^\top \mathbf{X}$. The data is now whitened because the scatter matrix of $\widetilde{\mathbf{X}}$ is the identity matrix $\mathbf{I}$. The whitened class means are now $\widetilde{\mathbf{m}}_k = \mathbf{P}^\top \mathbf{m}_k$. We have previously proven [13] that the generalized eigenvalue of the Fisher Criterion (see Equation (1)) is equal to the ratio $\frac{\lambda_b}{\lambda_w}$, where λ_b, λ_w are the eigenvalues of $\widetilde{\mathbf{S}}_b, \widetilde{\mathbf{S}}_w$ corresponding to the same eigenvector, where $\widetilde{\mathbf{S}}_b = \mathbf{P}^\top \mathbf{S}_b \mathbf{P}$ and $\widetilde{\mathbf{S}}_w = \mathbf{P}^\top \mathbf{S}_w \mathbf{P}$.

Suppose $\mathbf{V}$ is the set of eigenvectors of $\widetilde{\mathbf{S}}_b$. We can partition the columns of $\mathbf{V}$ as shown below, according to their corresponding eigenvalues λ_b: (1) those columns whose $\lambda_b = 1$; (2) those columns whose $0 < \lambda_b < 1$; and (3) those columns whose $\lambda_b = 0$, respectively. (We can easily calculate λ_w by noting that $\lambda_b + \lambda_w = 1$.)

$$\mathbf{V} = [\, \mathbf{V}_1 \mid \mathbf{V}_2 \mid \mathbf{V}_3 \,] \tag{4}$$

We will refer to the subspaces spanned by the three mutually orthogonal matrices as *Identity Space*, *Mixed Space*, and *Variation Space*, respectively.

3. Identity Space

We can now study the properties of the subspace spanned by $\mathbf{V}_1$, which we term *Identity Space*. We will show that (a) it contains only the class means; and (b) all within-class variation has been "projected out". That is, Identity Space reveals the class label (identity) of a data point, hence justifying its name. We begin with the following theorem.

Theorem 1. *In WFLD, if* $\mathbf{V}_1$ *is the set of eigenvectors of* $\widetilde{\mathbf{S}}_w$ *associated with* $\lambda_w = 0$*, then*

$$\mathbf{V}_1^\top \widetilde{\mathbf{x}}_i = \mathbf{V}_1^\top \widetilde{\mathbf{m}}_k, \qquad \forall \widetilde{\mathbf{x}}_i \in L_k \tag{5}$$

For detailed proof, please refer to our another paper [11]. A few remarks are in order. First, we see that the vector $\mathbf{m}'_k = \mathbf{V}_1^\top \widetilde{\mathbf{m}}_k \in \mathbb{R}^{(r_t - r_w) \times 1}$, may be used to represent the identity of class L_k, since all samples of the class project onto it. We may therefore call it the *Identity Vector*. Second, all within-class variation are projected out of Identity Space. Third, a sample from one class will never project onto the Identity Vector of another distinct class.

3.1. Geometric structure

Having established that Identity Space contains only identity information (*i.e*. between-class variation), we now ask about its structure: how the Identity Vectors are arranged in Identity Space. This can be answered by examining $\widetilde{\mathbf{S}}_b$, the dual of $\widetilde{\mathbf{S}}_w$, as shown in the next theorem:

Theorem 2. *Given the C Identity Vectors $\{\mathbf{m}'_k\}$, if the dimension of Identity Space is $C-1$, then*

$$\|\mathbf{m}'_k\| = \sqrt{\frac{1}{N_k} - \frac{1}{N}}, \tag{6}$$

$$\cos\theta_{kl} = \frac{-\sqrt{N_k N_l}}{\sqrt{N-N_k}\sqrt{N-N_l}}, \quad \forall\, k \neq l \tag{7}$$

where θ_{kl} is the angle between $\mathbf{m}'_k$ and $\mathbf{m}'_l$, N_k is the number of samples in class L_k, and $N = \sum N_k$.

Please see Appendix A for the proof.

Corollary 3. *If each class has the same number of samples, $\forall k,\ N_k = n$, then*

$$\|\mathbf{m}'_k\| = \sqrt{\frac{C-1}{nC}}, \quad \textit{and} \quad \cos\theta_{kl} = -\frac{1}{C-1}, \quad \forall\, k \neq l \tag{8}$$

We can easily prove Corollary 3 by setting $N_k = n$ in Theorem 2. Corollary 3 says that with equal samples in all classes, the C Identity Vectors distribute on a hypersphere of radius $r_C = \sqrt{\frac{C-1}{nC}}$. Moreover, the angle between any two Identity Vectors is constant, and always larger than $90°$. This means that the Identity Vectors are maximally spread out on the hypersphere. (They form the vertices of a regular simplex in $\mathbb{R}^{C-1}$.) We may therefore call this the *Identity Sphere*. Figures 1(a) and 1(b) show the Identity Spheres for $C = 3$ and 4.

4. Variation Space

We now study the properties of Variation Space, the subspace spanned by $\mathbf{V}_3$. We will prove that in this subspace, (a) all class means project to zero; and (b) the within-class variations of each class lie in orthogonal subspaces.

Theorem 4. *In WFLD, if $\mathbf{V}_3$ is the set of eigenvectors of $\widetilde{\mathbf{S}}_b$ associated with $\lambda_b = 0$, then all class means project to $\mathbf{0}$:*

$$\forall\, k, \qquad \mathbf{V}_3^\top \widetilde{\mathbf{m}}_k = \mathbf{0}. \tag{9}$$

Proof. For any eigenvector $\mathbf{v} \in \mathbf{V}_3$, we have $\mathbf{v}^\top \widetilde{\mathbf{S}}_b \mathbf{v} = 0$. But $\widetilde{\mathbf{S}}_b = \mathbf{P}^\top \mathbf{S}_b \mathbf{P}$, so $\mathbf{v}^\top \mathbf{P}^\top \mathbf{S}_b \mathbf{P}\mathbf{v} = 0$. If we replace $\mathbf{S}_b$ with Equation (2), then

$$0 = \sum_{k=1}^{C} N_k \mathbf{v}^\top \mathbf{P}^\top \mathbf{m}_k \mathbf{m}_k^\top \mathbf{P}\mathbf{v} \tag{10}$$

$$= \sum_{k=1}^{C} N_k \|\mathbf{v}^\top \widetilde{\mathbf{m}}_k\|^2 \tag{11}$$

This is a sum of squared norms, which is zero if and only if each term is zero. Hence $\mathbf{v}^\top \widetilde{\mathbf{m}}_k = 0$. But this is true for any $\mathbf{v} \in \mathbf{V}_3$, and thus $\mathbf{V}_3^\top \widetilde{\mathbf{m}}_k = \mathbf{0}$. □

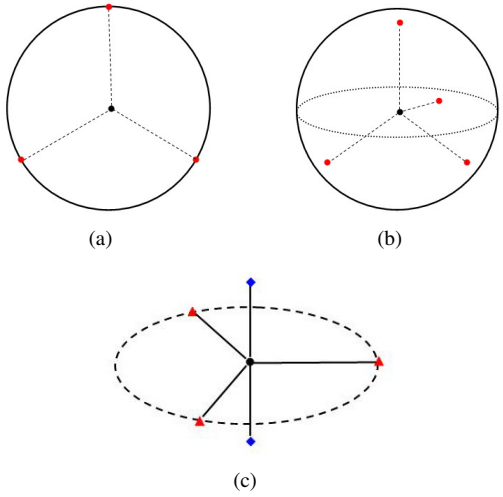


Figure 1. (a) — (c) Illustration of the Identity Sphere and Variation Space. (a) When $C = 3$, it is a circle in 2D space and the Identity Vectors are vertices of an equilateral triangle; (b) When $C = 4$, it is a sphere in 3D space and the Identity Vectors are vertices of a regular tetrahedron; (c) An illustration of the geometric structure of Variation Space. Note that different shapes (colors) represent distinct classes. Class One has two samples lying on a vertical line, and both are orthogonal to Class Two, which has three samples evenly distributed on a circle.

4.1. Geometric structure

Let's explore the geometric structure of Variation Space.

Lemma 5. *The inner product of any two whitened samples $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_j$ is:*

$$\widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j = \begin{cases} 1 - \frac{1}{N} & \textit{if} \quad i = j, \\ -\frac{1}{N} & \textit{if} \quad i \neq j. \end{cases} \tag{12}$$

Please refer to Appendix B for a proof.

Theorem 6. *After projecting onto Variation Space, any two vectors $\mathbf{V}_3^\top \widetilde{\mathbf{x}}_i = \mathbf{x}'_i \in L_k$ and $\mathbf{V}_3^\top \widetilde{\mathbf{x}}_j = \mathbf{x}'_j \in L_l$, have their inner product given by:*

$$\mathbf{x}_i'^\top \mathbf{x}'_j = \begin{cases} 1 - \frac{1}{N_k} & \textit{if} \quad i = j \textit{ and } L_k = L_l, \\ -\frac{1}{N_k} & \textit{if} \quad i \neq j \textit{ and } L_k = L_l, \\ 0 & \textit{if} \quad i \neq j \textit{ and } L_k \neq L_l. \end{cases} \tag{13}$$

Again, we defer the proof to Appendix C so as not to distract from the main flow of our paper. Theorem 6 says that within Variation Space, any two classes lie in orthogonal subspaces. This can be seen from the last clause of Equation (13). Because of this, the within-class variation of each class do not overlap, but instead occupy distinct regions in the subspace $\mathbb{R}^{r_t - r_b}$. Moreover, the Theorem also reveals how class samples are distributed in Variation Space, as shown in the next Corollary.

Corollary 7. *After projecting onto Variation Space, all vectors from the same class L_k have the same length $\sqrt{1-\frac{1}{N_k}}$; and any two vectors are separated by a constant angle θ_k, where $\cos\theta_k = \frac{-1}{N_k-1}$.*

Here then is the whole picture of Variation Space. (1) Different classes share the same mean (Theorem 4); (2) Any two classes are orthogonal to each other (Theorem 6); (3) The N_k vectors for each class are equally distributed over a hypersphere of dimension $N_k - 1$ (Corollary 7). That is, the N_k vectors form a regular simplex. Figure 1(c) shows an example of Variation Space for two classes $N_1 = 2$ and $N_2 = 3$.

5. Practical Considerations

5.1. Existence conditions

The Identity and Variation Spaces do not always exist; their existence depends on r_b, r_t, and r_w respectively. For the Identity Space to exist at its maximum extent (size = $C-1$), a sufficient condition is that (a) all data samples are linearly independent, and (b) $D \geq N-1$. This is also the sufficient condition for Variation Space to exist at its maximum size of $N-C$. Moreover, in this case, the size of Mixed Space is zero. This is the ideal situation because all class samples are neatly separated into Identity Space and Variation Space. Equation (4) now becomes: $\mathbf{V} = [\,\mathbf{V}_1 \mid \mathbf{V}_3\,]$. Note that V is an $(N-1)\times(N-1)$ orthogonal matrix, and thus invertible.

In practice, Identity or Variation Space does not always exist because the sufficient condition ($D \geq N-1$ and linear independence) may not be satisfied. To encourage their existence, we implicitly transform the data into a high-dimensional space by using the kernel trick [1]. There are two reasons to employ the kernel mapping function φ. First, the function φ can map two linearly dependent vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ onto two linearly independent ones $\varphi(\mathbf{x}_i)$ and $\varphi(\mathbf{x}_j)$ [4]. Second, the mapped space $\{\varphi(\mathbf{x}_i)\}$ could have arbitrarily large (even infinite) dimensionality [9]. After mapping the original space into the high-dimensional space with some function φ, we then maximize the Fisher Criterion in the new transformed space. This method is termed kernel WFLD (kWFLD) hereafter.

5.2. Decomposition and Representation

We can thus decompose any training point $\widetilde{\mathbf{x}}_i \in L_k$ into two components:

$$\begin{aligned}\widetilde{\mathbf{x}}_i &= \mathbf{V}_1\mathbf{V}_1^\top\widetilde{\mathbf{x}}_i + \mathbf{V}_3\mathbf{V}_3^\top\widetilde{\mathbf{x}}_i &(14)\\ &= \mathbf{V}_1\mathbf{V}_1^\top\widetilde{\mathbf{m}}_k + \mathbf{V}_3\mathbf{V}_3^\top\widetilde{\mathbf{x}}_i &(15)\\ &= \mathbf{V}_1\mathbf{m}'_k + \mathbf{V}_3\mathbf{x}'_i. &(16)\end{aligned}$$

where $\mathbf{x}'_i = \mathbf{V}_3^\top\widetilde{\mathbf{x}}_i$ is the projection onto Variation Space, and $\mathbf{m}'_k = \mathbf{V}_1^\top\widetilde{\mathbf{x}}_i = \mathbf{V}_1^\top\widetilde{\mathbf{m}}_k$ is the projection onto Identity Space. This decomposition follows because $\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_3\mathbf{V}_3^\top = \mathbf{I}$. Thus, any sample $\widetilde{\mathbf{x}}_i \in L_k$ can be decomposed into $\mathbf{m}'_k$ (the identity component), and $\mathbf{x}'_i$ (the variation component). Furthermore, $\mathbf{x}'_i \in \mathbb{R}^{N-C}$ has a sparse representation (only $N_k - 1$ nonzeros) because the within-class variation of each class is distinct (see Theorem 6). To recap, we have achieved a clean decomposition of a sample into its identity and variation components. This has the following merits.

1. It provides a way for efficient data representation. Any data point $\mathbf{x}_i \in \mathbb{R}^D$ can be decomposed into its identity vector $\mathbf{m}'_k \in \mathbb{R}^{C-1}$ and its variation vector $\mathbf{x}'_i$ encoded by $N_k - 1$ nonzero numbers, requiring only $C + N_k - 2$ numbers to represent it. Note that C and N could be $\ll D$, *e.g.* in face recognition, typically $D \approx 10^4$, while $C \approx 100$ and $N_k \approx 10$.

2. It combines the strengths of PCA (which is suited for representation, not classification) and the FLD (which is meant for classification, not representation). The identity vector $\mathbf{m}'_k$ is best for classification according to the Fisher Criterion, while the vector $\mathbf{x}'_i$ makes lossless reconstruction possible by retaining the variation information.

6. Experiments

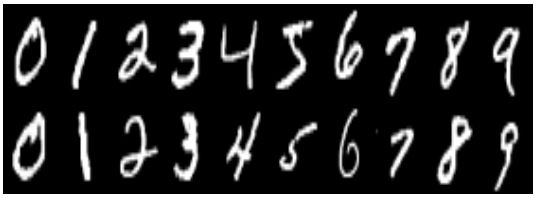
6.1. Discriminability of Identity Space

We perform face and digits recognition by using two datasets. The experimental setting is as follows. For PCA, we take the top C principal components, where C is the number of classes; for LDA, we apply PCA first by keeping 95% eigen-energy, followed by LDA; for kernel WFLD, we use Gaussian kernel. After projection, we use 1NN to perform classification. The recognition rate is reported by averaging over 20 runs.

1. Banca dataset [2] contains 52 subjects, and each subject has 120 face images, which are normalized to 51×55 in pixels. By using a web cam and an expensive camera, these subjects were recorded in three different scenarios over a period of three months. Each face image contains illumination, expression and pose variations because the subjects are required to talk during the recording (Fig. 2(a)).

2. MNIST dataset is derived from the NIST dataset, and has been created by Yann LeCun [8]. This dataset of handwritten digits ($'0' - '9'$) has a set of $70,000$ examples in total. The digits have been centered and normalized to 28×28 in pixels. Fig. 2(b) shows a sample of MNIST digits images.

(a)



(b)

(c)

Figure 2. Samples of real data: (a) Banca faces; (b) MNIST digits; (c) PIE faces. For (a) and (c), each row represents one person. Note that PIE dataset only presents illumination variation; whereas, Banca dataset presents more variations, such as illumination, pose, and expression.

For face recognition, we randomly choose n training samples from each subject, $n = 2, \cdots, 12$, and the remaining images are used for testing. For each set of n training samples, we employ cross validation so that we can compute the mean and standard deviation for classification accuracies. As shown in Table 1, we observe that for all methods, the more training samples, the greater the recognition accuracy. Kernel WFLD achieves around 2% better accuracy than WFLD. This shows that kernel method makes classes more separable in the high dimensional space. However, this performance difference is only slight, because Identity Space already exists in the original space for linearly independent samples with $D \geq N - 1$.

For digits recognition, we randomly choose $n = 100, \cdots, 600$ training samples from each class, and the remaining images are used for testing. As shown in Table 2, WFLD produces the worst performance among all four methods. The reason is that Identity Space does not exist, so that identity and variation information are mixed together in Mixed Space. Kernel WFLD is always the best classifier in terms of the classification accuracy; it is also the most stable classifier in terms of having the smallest standard deviation. The gap between kWFLD and the second best is around 10%. All these results demonstrate that even when Identity Space does not exist in the original data space when $N > D$, we can still create the most discriminant subspaces (aka Identity Space) by using the kernel method.

Another strength of WFLD and kWFLD is efficiency. To classify an unknown point, WFLD or kWFLD needs to compare it with C Identity Vectors, whereas PCA and LDA require nC comparisons. For digits recognition, $C = 10$ and $n = 600$, so WFLD is 600 times more efficient in terms of storage and computation.

6.2. Discriminability of Variation Space

We now compare the discriminative power in Identity Space and Variation Space. We do this by taking a weighted sum of d_k^V and d_k^I:

$$d_k = \alpha d_k^I + (1 - \alpha) d_k^V, \quad 0 \leq \alpha \leq 1. \tag{17}$$

where d_k^I is the distance from the query point to each Identity Vector and d_k^V is the distance from the query point to class Variation Space. The classification is performed by using the minimum distance: $L^\star = \arg\min_{L_k} d_k$. Obviously, when $\alpha = 1$, we are using only Identity Space, and when $\alpha = 0$, we are using only Variation Space. To illustrate, we will use the Banca dataset for face recognition under varying illumination, pose, and expression.

The experimental setting is thus: From the Banca dataset, we choose $n = 2, 4, 6, 8, 10, 12$ samples per class as the training set, and the rest for testing. We then vary $\alpha \in [0, 1]$ in steps of 0.1, and in each case classify the testing samples according to minimum distance. We repeat the sampling 20 times to compute the mean and standard deviation of classification accuracy. Figure 3 shows the results [1]. We observe that:

1. Variation Space does contain some discriminative information. This can be seen in the row where $\alpha = 0$: the accuracies are significantly better than random guessing ($= \frac{1}{52} = 1.92\%$).

2. Identity Space is more discriminative than Variation Space, by comparing the accuracies for $\alpha = 0$ and $\alpha = 1$. More precisely, Identity Space achieves at least 33% greater accuracy than Variation Space.

3. Variation and Identity Spaces overlap a lot in terms of discriminability. This is suggested by the fact that when $\alpha = 1$, the performance is approximately the best. This substantiates the effectiveness of Identity Space as a pattern recognition tool.

[1] For the purpose of clear visualization, we only plot the mean rates without the standard deviation.

Table 1. BANCA: Classification accuracy (%) with different training set size.

#.	2	4	6	8	10	12
PCA	38.30 (1.57)	54.45 (1.78)	65.37 (0.74)	73.63 (1.04)	78.54 (1.33)	82.34 (1.31)
LDA	65.33 (3.71)	79.81 (2.19)	90.15 (0.94)	94.20 (0.85)	95.93 (0.79)	96.92 (0.48)
WFLD	**70.45** (1.74)	85.33 (1.67)	90.87 (0.77)	93.85 (0.56)	94.35 (0.60)	95.12 (0.35)
kWFLD	70.26 (1.74)	**86.89** (2.12)	**92.29** (1.21)	**95.44** (0.59)	**96.47** (0.56)	**97.04** (0.41)

Table 2. MNIST: Classification accuracy (%) with different training set size.

#.	100	200	300	400	500	600
PCA	82.76 (0.57)	85.23 (0.50)	86.49 (0.34)	87.13 (0.27)	87.75 (0.21)	88.14 (0.24)
LDA	82.69 (0.53)	85.12 (0.37)	86.10 (0.29)	86.72 (0.23)	87.24 (0.21)	87.50 (0.23)
WFLD	63.38 (1.03)	76.41 (0.49)	79.72 (0.34)	81.37 (0.29)	82.26 (0.27)	82.84 (0.25)
kWFLD	**92.79** (0.24)	**94.84** (0.12)	**95.70** (0.11)	**96.19** (0.11)	**96.54** (0.08)	**96.79** (0.08)

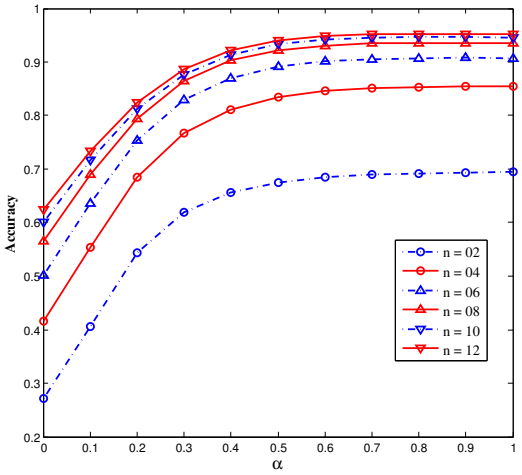


Figure 3. Face recognition by using both Identity and Variation Spaces with varying training set size $n = 2, ..., 12$. When $\alpha = 1$, we are using only Identity Space; when $\alpha = 0$, we are using only Variation Space. It shows that Identity Space contains much more discriminant information than Variation Space.

6.3. Face Synthesis

Our next experiment attempts to synthesize face images under varying illumination. We know from Section 5.2 that we can use WFLD to cleanly decompose any sample into its identity and variation components, modify each component separately, and then reconstruct a new sample. More specifically, given a face image $\widetilde{\mathbf{x}}_i$, we synthesize a new image using $\widehat{\mathbf{x}}(c) = \mathbf{V}_1\mathbf{V}_1^\top\widetilde{\mathbf{x}}_i + c\mathbf{V}_3\mathbf{V}_3^\top\widetilde{\mathbf{x}}_i$, where c is a parameter for us to control the amount of variation.

To illustrate, we will use the CMU PIE [10] dataset for illumination synthesis. We choose $C = 68$ subjects, each having 24 frontal face images taken under a wide range of lighting conditions. All face images are aligned based on eye coordinates, and cropped to the size of $70 \times 80 (= D)$ pixels. Figure 2(c) shows some PIE face images.

Figure 4 shows two examples of illumination synthesis as we vary $c = -1, \ldots, 3$. It is evident that this parameter controls the illumination. Because the WFLD is a lossless invertible transform, when $c = 1$ we get back exactly the input image. Here, it is clear from the shadows that the light source is from the left of the face. When $c = 0$, the illumination appears frontal and there is no shadow. This is because in the PIE dataset, the illumination varies approximately symmetrically from left to right, so that the mean face is frontally illuminated. Since $c = 0$, we are effectively reconstructing the face using only its identity component and suppressing its variation component. This appears to have suppressed shadows as well. When $c = -1$, we observe that the illumination has moved to the right; while for $c = 2, 3$, the illumination has become harsh.

Actually, we have a whole subspace (the class variation space of the sample $\widetilde{\mathbf{x}}_i$) to vary the illumination. Any vector in this subspace will synthesize a new illumination. We have merely shown the trivial case of varying the magnitude of the projected sample.

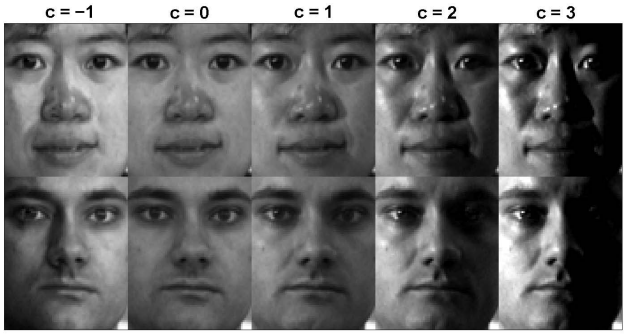
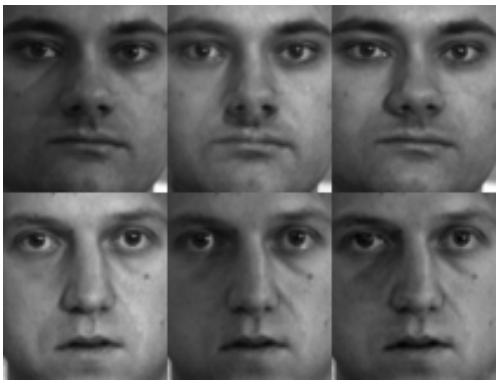


Figure 4. Two examples of illumination synthesis, as c, the user-controlled parameter is varied. The original input images are in column $c = 1$, because the WFLD is an invertible transform.

Our final experiment shows that WFLD may be used to swap face illuminations. Fig. 5 shows two examples of the variation synthesis. Given two novel face images under different lighting, we synthesize the face images with the swapped variation components. In this case, the variation components represent the face illumination. To evaluate the

synthesis quality, we also show the real face images under the swapped lighting. Our synthesized face images are comparable to the real face images, except the shadows. The reason is the shadows are generally encoded as high-order information. However, our theory is based on second-order information (scatter matrices).



(a)

(b)

Figure 5. Two examples of variation synthesis. For (a) and (b), the leftmost column is the given face images with different identities and variations (lighting); the medium column shows the synthesized face images with the swapped lighting; the rightmost column shows the real face images with the swapped lighting.

7. Conclusion

We have proven a number of important theoretical properties of the WFLD. As such, we argue that the WFLD is the correct way to compute the FLD. In so far as the original objective of the FLD was to eliminate within-class variation and maximize between-class variation, the pre-whitening step guarantees this objective. Note that we are *not* claiming that the WFLD is better than the Bayes' Classifier. Instead, we are merely explaining that elegant properties emerge when the FLD is computed correctly.

More precisely, the WFLD decomposes pattern vectors into two neat subspaces with useful properties: Identity and Variation Spaces. We defined Identity Space as the optimal subspace for pattern classification in terms of the Fisher Criterion, and then mathematically proved its property: all samples from one class project onto its Identity Vector, which is also its class mean. Variation Space was defined as the least discriminant subspace for pattern classification, and it contains no class means to be used for classification. We further showed their geometric structures, the conditions required for them to exist. When the conditions are not satisfied, the kernel trick was proposed to encourage the existence. To assess the performance of these two spaces, we ran face and digits recognition by using the Identity Space, and synthesized face images by using the Variation Space. The results show the power of WFLD as a pattern classification tool and as a data decomposition tool.

One limitation of our work is generalization. The property of clean class separation that we prove applies *only* to the training data, and not necessarily to novel, unseen data. Pre-whitening the FLD does not guarantee improved generalization, so standard regularization techniques [6] can still be applied here. However, it is clear from the linearity of our proofs that if novel data lie within the subspace of the training data, then they will also be perfectly separated. Thus, the crux is whether training data is representative of novel data — a problem common to all machine learning methods, not just to the FLD.

A. Proof of Theorem 2

Proof. The whitened $\mathbf{S}_b$ is decomposed as $\widetilde{\mathbf{S}}_b = \mathbf{V}\mathbf{\Lambda}_b\mathbf{V}^\top$. Within Identity Space, $\lambda_b = 1$. That is: $\mathbf{\Lambda}_b = \mathbf{I}$ corresponding to the set of eigenvectors $\mathbf{V}_1$. Thus, $\mathbf{V}_1^\top\widetilde{\mathbf{S}}_b\mathbf{V}_1 = \mathbf{I}$. Because $\widetilde{\mathbf{S}}_b = \mathbf{P}^\top\mathbf{S}_b\mathbf{P}$, it is easy to see that $\mathbf{V}_1^\top\mathbf{P}^\top\mathbf{S}_b\mathbf{P}\mathbf{V}_1 = \mathbf{I}$. If we replace $\mathbf{S}_b$ with Equation (2),

$$\mathbf{I} = \mathbf{V}_1^\top\mathbf{P}^\top\mathbf{S}_b\mathbf{P}\mathbf{V}_1 = \sum_{k=1}^{C} N_k\mathbf{V}_1^\top\mathbf{P}^\top\mathbf{m}_k\mathbf{m}_k^\top\mathbf{P}\mathbf{V}_1 \quad (18)$$

$$= \sum_{k=1}^{C} N_k\mathbf{m}'_k\mathbf{m}'^\top_k \quad (19)$$

Now let's denote $\mathbf{M} = [\sqrt{N_1}\mathbf{m}'_1, \cdots, \sqrt{N_C}\mathbf{m}'_C] \in \mathbb{R}^{(C-1)\times C}$, then we can rewrite Equation (19) as

$$\mathbf{M}\mathbf{M}^\top = \mathbf{I}_{C-1}. \quad (20)$$

where $\mathbf{I}_{C-1}$ is the $(C-1)\times(C-1)$ identity matrix. Since the total mean is zero, *i.e.* $\sum N_k\mathbf{m}'_k = \mathbf{0}$, it is easy to see that

$$\mathbf{M}\mathbf{i} = \mathbf{0} \quad (21)$$

where $\mathbf{i} = [\sqrt{N_1}, \cdots, \sqrt{N_C}]^\top$. Equations (20) and (21) shows that the rank of $\mathbf{M}$ is $C-1$ and its only nullspace is $\mathbf{i} \in \mathbb{R}^C$.

To compute $\mathbf{M}^\top\mathbf{M}$, we first define $\mathbf{Q} = \begin{bmatrix} \mathbf{M} \\ \frac{1}{\sqrt{N}}\mathbf{i}^\top \end{bmatrix} \in \mathbb{R}^{C\times C}$. Thus,

$$\mathbf{Q}\mathbf{Q}^\top = \begin{bmatrix} \mathbf{M} \\ \frac{1}{\sqrt{N}}\mathbf{i}^\top \end{bmatrix}\begin{bmatrix} \mathbf{M}^\top & \frac{1}{\sqrt{N}}\mathbf{i} \end{bmatrix} = \mathbf{I}_C \quad (22)$$

Here $N = \mathbf{i}^\top\mathbf{i} = \sum N_k$. Since $\mathbf{Q}$ is a full rank, square matrix, Equation (22) shows that $\mathbf{Q}$ is an orthogonal matrix, *i.e.* $\mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$. Hence,

$$\mathbf{Q}^\top\mathbf{Q} = \mathbf{M}^\top\mathbf{M} + \frac{1}{N}\mathbf{i}\mathbf{i}^\top = \mathbf{I} \quad (23)$$

$$\text{And thus,} \quad \mathbf{M}^\top \mathbf{M} = \mathbf{I} - \frac{1}{N}\mathbf{i}\mathbf{i}^\top \tag{24}$$

On the other hand,

$$\mathbf{M}^\top \mathbf{M} = \left[\sqrt{N_k N_l}\mathbf{m}_k'^\top \mathbf{m}_l'\right], \mathbf{i}\mathbf{i}^\top = \left[\sqrt{N_k N_l}\right]. \tag{25}$$

Considering Equations. (24) and (25), we see that

$$\sqrt{N_k N_l}\mathbf{m}_k'^\top \mathbf{m}_l' = \begin{cases} 1 - \frac{\sqrt{N_k N_l}}{N} & \text{if } k = l; \\ -\frac{\sqrt{N_k N_l}}{N} & \text{if } k \neq l. \end{cases} \tag{26}$$

$$\text{That is:} \quad \mathbf{m}_k'^\top \mathbf{m}_l' = \begin{cases} \frac{1}{N_k} - \frac{1}{N} & \text{if } k = l; \\ -\frac{1}{N} & \text{if } k \neq l. \end{cases} \tag{27}$$

$$\text{When } k = l, \qquad \|\mathbf{m}_k'\| = \sqrt{\frac{1}{N_k} - \frac{1}{N}}. \tag{28}$$

$$\text{When } k \neq l, \qquad \cos\theta_{kl} = \frac{\mathbf{m}_k'^\top \mathbf{m}_l'}{\|\mathbf{m}_k'\|\|\mathbf{m}_l'\|} = \frac{-\sqrt{N_k N_l}}{\sqrt{N - N_k}\sqrt{N - N_l}} \tag{29}$$

This completes the proof. □

B. Proof of Lemma 5

Proof. After pre-whitening, $\mathbf{P}^\top \mathbf{S}_t \mathbf{P} = \mathbf{I}$. That is:

$$\mathbf{I} = \mathbf{P}^\top \sum \mathbf{x}_i \mathbf{x}_i^\top \mathbf{P} = \sum \widetilde{\mathbf{x}}_i \widetilde{\mathbf{x}}_i^\top \tag{30}$$

We further define $\widetilde{\mathbf{X}} = [\widetilde{\mathbf{x}}_1, \cdots, \widetilde{\mathbf{x}}_N]$, then $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^\top = \mathbf{I}$. Because of the zero global mean, $\widetilde{\mathbf{X}}\mathbf{1} = \mathbf{0}$. Thus, $\text{rank}(\widetilde{\mathbf{X}}) = N - 1$ and its only nullspace is $\mathbf{1} \in \mathbb{R}^N$. Using the same trick as in Equations (22) to (24), we get

$$\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \tag{31}$$

The diagonal entries of $\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}$ are $1 - \frac{1}{N}$, and off-diagonal ones are $-\frac{1}{N}$. This completes the proof. □

C. Proof of Theorem 6

Proof.

$$\begin{aligned}
\mathbf{x}_i'^\top \mathbf{x}_j' &= \left(\mathbf{V}_3^\top \widetilde{\mathbf{x}}_i\right)^\top \left(\mathbf{V}_3^\top \widetilde{\mathbf{x}}_j\right) && (32)\\
&= \widetilde{\mathbf{x}}_i^\top \mathbf{V}_3 \mathbf{V}_3^\top \widetilde{\mathbf{x}}_j \quad = \widetilde{\mathbf{x}}_i^\top \left(\mathbf{I} - \mathbf{V}_1 \mathbf{V}_1^\top\right) \widetilde{\mathbf{x}}_j && (33)\\
&= \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j - \left(\mathbf{V}_1^\top \widetilde{\mathbf{x}}_i\right)^\top \left(\mathbf{V}_1^\top \widetilde{\mathbf{x}}_j\right) && (34)\\
&= \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j - \left(\mathbf{V}_1^\top \widetilde{\mathbf{m}}_k\right)^\top \mathbf{V}_1^\top \widetilde{\mathbf{m}}_l && (35)\\
&= \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j - \mathbf{m}_k'^\top \mathbf{m}_l' && (36)
\end{aligned}$$

Remarks: (1) Equation (33) is derived from $\mathbf{V}_1\mathbf{V}_1^\top + \mathbf{V}_3\mathbf{V}_3^\top = \mathbf{I}$. (2) Equation (36) is derived from Theorem 1. It remains to consider three cases, with the help of Equation (27) and Lemma 5:

1. When $i = j$, immediately $L_k = L_l$, then

$$\begin{aligned}
\mathbf{x}_i'^\top \mathbf{x}_j' &= \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j - \mathbf{m}_k'^\top \mathbf{m}_l' && (37)\\
&= \left(1 - \frac{1}{N}\right) - \left(\frac{1}{N_k} - \frac{1}{N}\right) \quad = 1 - \frac{1}{N_k} && (38)
\end{aligned}$$

2. When $i \neq j$ and $L_k = L_l$, then

$$\begin{aligned}
\mathbf{x}_i'^\top \mathbf{x}_j' &= \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j - \mathbf{m}_k'^\top \mathbf{m}_l' && (39)\\
&= -\frac{1}{N} - \left(\frac{1}{N_k} - \frac{1}{N}\right) \quad = -\frac{1}{N_k} && (40)
\end{aligned}$$

3. When $i \neq j$ and $L_k \neq L_l$, then

$$\begin{aligned}
\mathbf{x}_i'^\top \mathbf{x}_j' &= \widetilde{\mathbf{x}}_i^\top \widetilde{\mathbf{x}}_j - \mathbf{m}_k'^\top \mathbf{m}_l' && (41)\\
&= -\frac{1}{N} - \left(-\frac{1}{N}\right) \quad = 0 && (42)
\end{aligned}$$

This completes the proof. □

References

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruíz, and J. P. Thiran. The BANCA Database and Evaluation Protocol. In *The 4th International Conference on Audio- and Video-based Biometric Person Authentication*, pages 625–638, 2003.

[3] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 1997.

[4] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification, 2nd Edition*. John Wiley and Sons, 2000.

[6] J. Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd Edition*. Academic Press, 1990.

[8] Y. LeCun, L. Bottou, Y. Benjio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[9] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press; 1st Edition, 2001.

[10] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

[11] T. Sim, S. Zhang, J. Li, and Y. Chen. Simultaneous and Orthogonal Decomposition of Data using Multimodal Discriminant Analysis. In *IEEE International Conference on Computer Vision*, 2009.

[12] H. Yu and H. Yang. A Direct LDA Algorithm for High-Dimensional Data - with Application to Face Recognition. *Pattern Recognition*, 34(10):2067–2070, 2001.

[13] S. Zhang and T. Sim. When Fisher Meets Fukunaga-Koontz: A New Look at Linear Discriminants. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 323–329, June 2006.

[14] S. Zhang and T. Sim. Discriminant Subspace Analysis: A Fukunaga-Koontz Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1732–1745, October 2007.