

Analysis of Chernoff Criterion for Linear Dimensionality Reduction

Jing Peng and Stefan Robila

Department of Computer Science
Montclair State University
Montclair, NJ 07043

{jing.peng,stefan.robila@montclair.edu}

Wei Fan

Exploratory Stream Process Group
IBM T.J.Watson Research
Hawthorne, NY 10532
weifan@us.ibm.com

Guna Seetharaman

Information Directorate
AFMC AFRL/RITB
Rome, New York
Gunasekaran.Seetharaman@rl.af.mil

Abstract—Well known linear discriminant analysis (LDA) based on the Fisher criterion is incapable of dealing with heteroscedasticity in data. However, in many practical applications we often encounter heteroscedastic data, i.e., within class scatter matrices can not be expected to be equal. A technique based on the Chernoff criterion for linear dimensionality reduction has been proposed recently. The technique extends well-known Fisher's LDA and is capable of exploiting information about heteroscedasticity in the data. While the Chernoff criterion has been shown to outperform the Fisher's, a clear understanding of its exact behavior is lacking. In addition, the criterion, as introduced, is rather complex, thereby making it difficult to clearly state its relationship to other linear dimensionality techniques. In this paper, we show precisely what can be expected from the Chernoff criterion and its relations to the Fisher criterion and Fukunaga-Koontz transform. Furthermore, we show that a recently proposed decomposition of the data space into four subspaces is incomplete. We provide arguments on how to best enrich the decomposition of the data space in order to account for heteroscedasticity in the data.

Index Terms—Chernoff distance; Dimensionality reduction; Linear discriminant analysis

I. INTRODUCTION

In classification, a large number of features or attributes often make the design of a classifier difficult and degrade its performance. This is particularly pronounced when the number of examples is small relative to the number of features. This fact is due to the curse of dimensionality. It states in simple terms that the number of examples required to properly compute a classifier grows exponentially with the number of features. For example, assuming features are correlated, approximating a binary distribution in a n dimensional feature space requires estimating $O(2^n)$ unknown variables [1]. In such situations, the problem often becomes intractable. This calls for reducing the number of features in constructing classifiers.

There are many dimensionality reduction techniques in the literature. The two most popular ones are principal components analysis (PCA) [2] and linear discriminant analysis (LDA) [3]. Both techniques have been successfully applied to a wide variety of practical problems [4], [5], [6], [7], [8], [9], [10], [11], [12]. By projecting data onto a linear subspace spanned by principal components, PCA achieves dimension reduction with the minimal data reconstruction error. On the other hand, without taking into account class information PCA cannot

compute discriminant information required by classifiers. In this paper, we are concerned with LDA.

In LDA, we are given a set of l examples:

$$z = \{(x_i, y_i)\}_{i=1}^l. \quad (1)$$

These examples are independently and identically distributed (i.i.d.) from the probability space $Z = X \times Y$. Here probability measure ρ is defined but unknown, $x_i \in X \subset \mathbb{R}^q$ are the q -dimensional inputs, and $y_i \in Y = [-M, M] \subset \mathbb{R}$ are scalar labels. According to Fisher's criterion, one has to find a projection matrix $W \in \mathbb{R}^{q \times d}$ that maximizes:

$$J_F(W) = \text{tr}(W^t S_w W)^{-1} W^t S_b W \quad (2)$$

where

$$S_b = \sum_{c=1}^C p_c (m_c - m)(m_c - m)^t$$

and

$$S_w = \sum_{c=1}^C p_c \sum_{i=1, x_i \in c}^{n_c} (x_i - m_c)(x_i - m_c)^t$$

are so-called between-class and within-class matrices. Here m represents the overall mean, m_c denotes the mean of class c , d denotes the dimensions of the reduced space, and t represents the transpose operator. Determining linear discriminants with the Fisher criterion is relatively efficient computationally.

While LDA based on the Fisher criterion simply tries to separate class means as much as possible, it is incapable of exploiting potential discriminant information that might exist in data in terms of the difference between within class matrices. That is, it can not explicitly handle heteroscedastic data, where the data do not have equal within class matrices. This limitation becomes more pronounced in the two-class case, where a reduction to only one dimension is sufficient [3]. In the multi-class case, a reduction to the number of dimensions that is no more than the number of classes.

For the multi-class case with C classes, linear reduction to $C - 1$ dimensions does not guarantee to capture all the relevant information for a classification task. Even if the $C - 1$ dimensions capture all discriminants, it is unclear how LDA based on the Fisher criterion will exploit them. To address this problem, a new criterion for linear dimensionality

reduction for the two class case, called the Chernoff criterion, is proposed that extends and improves upon the Fisher criterion by taking the heteroscedasticity of the data into account [13]. The technique makes use of directed distance matrices (DDMs) [14], which can be viewed as a generalization of the between-class matrix. It is argued that the between-class matrix can be associated with squared Euclidean distance between pairs of class means.

While the Chernoff criterion is shown to outperform the Fisher criterion, a clear understanding of its exact behavior is lacking. In addition, the criterion, as introduced, is rather complex (especially in the multi-class case), thereby making it difficult to clearly state its relationship to other linear dimensionality techniques in general, and the Fisher criterion in particular. In this paper, we show precisely what can be expected from the Chernoff criterion and its relations to the Fisher criterion. Furthermore, we show that the decomposition of the data space into four subspaces described in [15] is incomplete. We demonstrate how to enrich the decomposition of the data space to account for heteroscedasticity in the data. In the current work, we focus on two class problems. This can be justified by the fact that Chernoff distance is intended for two distributions.

The paper is organized as follows. Section 2 discusses related work in this area. Section 3 formally defines the Chernoff criterion. Section 4 provides an analysis on its properties, and points out its relationship to the Fisher criterion and Fukunaga-Koontz transform. Section 5 provides an argument on how to best characterize the data space in order to account for heteroscedasticity in data. Section 6 presents some simple experimental evaluation. Finally, Section 7 summaries our work and points out future research.

II. RELATED WORK

Several methods for extending linear classifiers to unequal covariance matrices and non Gaussian distributions are discussed in [3]. These methods can be applied to dimensionality reduction as well. However, most of these techniques are derived for the two-class case and not readily extendable to the multi-class case. In addition, many require iterative optimization.

A technique based on Kullback divergence to extend the Fisher criterion is proposed in [16]. Several techniques based on probabilistic separability and interclass distance measures can be introduced in [17]. These techniques require rather time-consuming iterative procedures to compute linear discriminants.

A maximum-likelihood approach to LDA is described in [18]. This technique generalizes the Fisher criterion in that it does not make the assumption that all classes have equal within class matrices. It iteratively maximizes a likelihood model.

Recently, a dimension reduction technique, called linear feature extraction (LFE) is introduced in [19]. Let x be an instance. We define the *near hit* or nh of x as its nearest neighbor that comes from the same class as x . Similarly, we define the *near miss* or nm as the nearest neighbor of x that

comes from the opposite class. Then the hypothesis margin of x with respect to labeled data L is defined as [20]

$$\sigma(x) = \|x - nm(x)\| - \|x - nh(x)\|. \quad (3)$$

The hypothesis margin is easy to compute and lower bounds the sample margin [20].

Let $h(x) = x - nh(x)$ and $m(x) = x - nm(x)$. We define two matrices, near hit S_h and near miss S_m , as follows.

$$S_h = \sum_{i=1}^l h(x_i)h(x_i)^t$$

and

$$S_m = \sum_{i=1}^l m(x_i)m(x_i)^t.$$

Instead of optimizing the margin (3) by selecting features, a technique described in [19] computes a linear transform that optimizes the following

$$\begin{aligned} \max_x \quad & x^t(S_m - S_h)x \\ & (x^t x)^2 = 1. \end{aligned} \quad (4)$$

To be less sensitive to noise, k near misses and hits are often used in practice to optimize the margin for some integer k [19].

If we allow near hits and near misses to be extended to the entire neighborhood,

$$\begin{aligned} S_m - S_h &= l(m_{+1}m_{+1}^t - 2m_{+1}m_{-1}^t + m_{-1}m_{-1}^t) \\ &= lS_b. \end{aligned} \quad (5)$$

Thus, maximizing the margin reduces to maximizing the between-class scatter matrix. Because it ignores the within-class scatter matrix, it cannot be optimal. This lends theoretical support to the practical observation that the average neighborhood for near hit and near miss in Relief should be somewhere between 1 and $l/2$ [19].

A metric space dimension reduction technique, called discriminant neighborhood embedding (NDE), is introduced in [21]. The idea is to find a linear transform such that in the transformed space total within class distance is minimized, while total between class distance is maximized. Let $x_j \in NB_w(x_i)$ if x_j is a within class neighbor of x_i , and $x_j \in NB_b(x_i)$ if x_j is a between class neighbor of x_i . The neighborhood can be computed using k nearest neighbors. Then this objective is achieved by defining an adjacency matrix F , where

$$F_{ij} = \begin{cases} 1 & \text{if } x_i \in NB_w(x_j) \text{ or } x_j \in NB_w(x_i); \\ -1 & \text{if } x_i \in NB_b(x_j) \text{ or } x_j \in NB_b(x_i); \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to find P such that

$$tr(P^t X(S - F)X^t P)$$

is minimized, subject to $P^t P = I$. Here X is the data matrix, and S is a diagonal matrix, where $S_{ii} = \sum_j F_{ij}$ or $S_{ii} = \sum_j F_{ji}$.

If we use $k = 1$ to compute NB_w and NB_b , we can write $X(S - F)$ as

$$M = (mh(x_1) \cdots mh(x_l)),$$

where $mh(x_i) = nm(x_i) - nh(x_i)$ represents the difference between the near miss $nm(x_i)$ and the near hit $nh(x_i)$ of x_i , respectively [20]. In this case, the objective becomes maximizing $tr(P^t M X^t P)$ over P , subject to $P^t P = I$. Here the local information is the cross covariance of an instance and the difference between its near miss and near hit. If we compute NB_w and NB_b over the entire classes, i.e., $k = l/2$ (assuming each class has the same number of examples), it can be shown that

$$X(S - F)X^t = S_b.$$

The result is similar to LFE (Eq. (5)). In practice, k is chosen somewhere between 1 and $l/2$.

A dimensionality technique that improves the Fisher criterion in heteroscedastic data has been proposed [13]. The technique employs the Chernoff distance to measure differences in both between and within class covariance. Like the Fisher criterion, it involves computing the inverse of within class matrices. Thus, it potentially suffers from the small sample size problem. A related technique that maximizes the Chernoff distance in the transformed space, thereby augmenting class separability in the space is introduced in [22]. Convergence analysis is also provided. More recently, an algorithm is proposed that computes the one-dimensional subspace where the Bayes error is minimized for multi-class problems with homoscedastic data distributions [23].

III. CHERNOFF CRITERION

The Fisher criterion (Eq. (2)) states that in order to compute linear discriminants LDA maximizes the ratio of the between class matrix to the average within class matrix in a reduced space. This is achieved by solving a generalized eigenvalue problem $S_b w = \lambda S_w w$ [3].

For the moment, we assume that the data is linearly transformed such that the S_w is identity. Then $J_F(W)$ can be maximized by taking the eigenvector v associated with the largest eigenvalue of $S_e = (m_1 - m_2)(m_1 - m_2)^t$. Note that $S_b = p_1 p_2 S_e$, where p_i is the a priori probability of class i . Notice that the eigenvalue equals the squared Euclidean distance. It can be shown that the matrix S_e provides us with the distance between two distributions, in addition to the direction (e.g., eigenvectors).

From the above, if discriminant information exists due to heteroscedasticity of the data, then this information should be present in DDMs. We note that this information about heteroscedasticity may not be in the same direction that separates class means. One powerful direct distance measure is based on the Chernoff distance that provides a measure between two probability density functions p_1 and p_2 :

$$D_C = -\log \int p_1^\alpha(x) p_1^{1-\alpha}(x) dx \quad (6)$$

where $\alpha \in \{0, 1\}$ is a constant. For two normally distributed densities, the Chernoff distance can be written as [24], [25]

$$D_C = (m_1 - m_2)^t (p_1 S_1 + p_2 S_2)^{-1} (m_1 - m_2) + \frac{1}{p_1 p_2} \log \frac{|p_1 S_1 + p_2 S_2|}{|S_1|^{p_1} |S_2|^{p_2}}. \quad (7)$$

It can be shown that one can obtain D_C as the trace of matrix S_C [13]

$$S_C = S^{-1/2} (m_1 - m_2) (m_1 - m_2)^t S^{-1/2} + \frac{1}{\alpha(1-\alpha)} (\log S - \alpha \log S_1 - (1-\alpha) \log S_2) \quad (8)$$

This provides a basis for the Chernoff criterion for linear dimensionality reduction.

For the moment, we assume that $S_w = I$. The Fisher criterion becomes:

$$J_F(W) = tr((W^t W)^{-1} (p_1 p_2 W^t S_e W)).$$

If we replace S_e by S_C , we obtain a heteroscedastic generalization of the Fisher Criterion. In general, when $S_w \neq I$, we can first transform the data by $S_w^{-1/2}$. In this space, the criterion for LDA becomes

$$tr((W^t W)^{-1} (p_1 p_2 W^t S_w^{-1/2} S_e S_w^{-1/2} W)).$$

We will then transform this back to the original space by $S_w^{1/2}$. For the Fisher criterion, we have [13]

$$tr((W^t S_w W)^{-1} (p_1 p_2 W^t S_e W)).$$

Now, replacing S_e by S_C , we arrive at the Chernoff criterion.

The heteroscedastic two-class Chernoff criterion J_C is defined as

$$J_C(W) = tr((W^t S_w W)^{-1} (W^t S_b W - W^t S_w^{\frac{1}{2}} (p_1 \log(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}) + p_2 \log(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}})) S_w^{\frac{1}{2}} W)) \quad (9)$$

IV. ANALYSIS OF THE CHERNOFF CRITERION

Here we examine the Chernoff criterion in detail by repeatedly applying the principle of simultaneous diagonalization of two matrices. This simultaneous diagonalization is based on Fukunaga-Koontz transform (FKT) [26]. Since $S_w = p_1 S_1 + p_2 S_2$, we can simultaneously diagonalize S_1 and S_2 . Let

$$P = Q \Lambda^{-\frac{1}{2}},$$

where $S_w = Q \Lambda Q^t$. Then

$$P^t S_w P = p_1 P^t S_1 P + p_2 P^t S_2 P = I.$$

Thus, it can be shown that $\tilde{S}_1 = P^t S_1 P$ and $\tilde{S}_2 = P^t S_2 P$ can be simultaneously diagonalized [3]

$$\tilde{S}_1 = V \Lambda^{(1)} V^t \quad (10)$$

and

$$\tilde{S}_2 = V \Lambda^{(2)} V^t \quad (11)$$

where $\Lambda^{(1)}$ and $\Lambda^{(2)}$ are the eigenvalue matrices of \tilde{S}_1 and \tilde{S}_2 , respectively, satisfying

$$p_1\Lambda^{(1)} + p_2\Lambda^{(2)} = I,$$

and V is the eigenvector matrix of both \tilde{S}_1 and \tilde{S}_2 , e.g., \tilde{S}_1 and \tilde{S}_2 share the same eigen space. In addition, the following conditions hold

$$V^t P^t S_w P V = I \quad \text{and} \quad V^t P^t S_1 P V = \Lambda^{(1)}$$

and

$$S_w^{-1} S_1 P V = P V \Lambda^{(1)}.$$

The above implies that

$$\begin{aligned} S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}} &= Q P^t S_1 P Q^t \\ &= Q \tilde{S}_1 Q^t \\ &= Q V \Lambda^{(1)} V^t Q^t \end{aligned} \quad (12)$$

Since V and Q are orthogonal, it follows that

$$\begin{aligned} p_1 \log(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}) &= p_1 \log(Q V \Lambda^{(1)} S_1 V^t Q^t) \\ &= p_1 Q V \log(\Lambda^{(1)}) V^t Q^t \\ &= p_1 Q V \tilde{\Lambda}^{(1)} V^t Q^t, \end{aligned} \quad (13)$$

where $\tilde{\Lambda}_i^{(1)} = \log(\lambda_i^{(1)})$. Similarly, we have

$$p_2 \log(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}}) = p_2 Q V \tilde{\Lambda}^{(2)} V^t Q^t, \quad (14)$$

where $\tilde{\Lambda}_i^{(2)} = \log(\lambda_i^{(2)})$. Let

$$\Sigma = p_1 \log(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}) + p_2 \log(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}}) \quad (15)$$

Combining Eqs. (13) and (14) gives rise to

$$\begin{aligned} \Sigma &= p_1 Q V \tilde{\Lambda}^{(1)} V^t Q^t + p_2 Q V \tilde{\Lambda}^{(2)} V^t Q^t \\ &= Q V (p_1 \tilde{\Lambda}^{(1)} + p_2 \tilde{\Lambda}^{(2)}) V^t Q^t \\ &= Q V \tilde{\Lambda} V^t Q^t, \end{aligned} \quad (16)$$

where

$$\tilde{\Lambda}_i = p_1 \tilde{\Lambda}_i^{(1)} + p_2 \tilde{\Lambda}_i^{(2)} \quad (17)$$

$$= \log((\lambda_i^{(1)})^{p_1} (\lambda_i^{(2)})^{p_2}) \quad (18)$$

Define

$$\tilde{S}_w = S_w^{\frac{1}{2}} \Sigma S_w^{\frac{1}{2}}. \quad (19)$$

Then the Chernoff criterion becomes

$$J_C(W) = \text{tr}((W^t S_w W)^{-1} (W^t (S_b - \tilde{S}_w) W)). \quad (20)$$

We can optimize $J_C(A)$ by solving an eigenvalue decomposition of the matrix $S_w^{-1} (S_b - \tilde{S}_w)$.

How different is the Chernoff criterion from Fisher's? We answer this question by examining the solution to Eq. (20), and thus Eq. (9). First, we write $S_w^{-1} (S_b - \tilde{S}_w)$ as $S_w^{-1} (S_b + \tilde{S}_w)$ by rewriting $\tilde{\Lambda}$ in Eq. (18)

$$\tilde{\Lambda}_i = \log\left(\frac{1}{(\lambda_i^{(1)})^{p_1} (\lambda_i^{(2)})^{p_2}}\right). \quad (21)$$

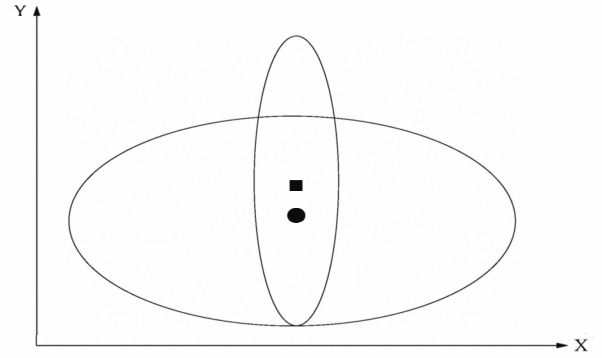


Fig. 1. In this example, the difference between the two co-variances is most pronounced along the X axis. Fisher chooses the Y axis as its discriminant, while Chernoff chooses the X axis.

We can simultaneously diagonalize S_w and $(S_b + \tilde{S}_w)$ using

$$X = Q \Lambda^{-\frac{1}{2}} U, \quad (22)$$

where Λ and Q are the eigenvalue and eigenvector matrices of S_w , and U diagonalizes

$$\Lambda^{-\frac{1}{2}} Q^t (S_b + \tilde{S}_w) Q \Lambda^{-\frac{1}{2}} = U D U^t, \quad (23)$$

where D is the eigenvalue matrix. That is,

$$X^t S_w X = I$$

and

$$X^t (S_b + \tilde{S}_w) X = D.$$

Furthermore, it can be shown that

$$S_w^{-1} (S_b + \tilde{S}_w) X = X D. \quad (24)$$

Thus, X is the eigenvector matrix of $S_w^{-1} (S_b + \tilde{S}_w)$ [3]. Eq. (24) can be simplified to

$$(\bar{S}_b + V \tilde{\Lambda} V^t) U = U D, \quad (25)$$

where

$$\bar{S}_b = \Lambda^{-\frac{1}{2}} Q^t S_b Q \Lambda^{-\frac{1}{2}}, \quad (26)$$

where $\bar{S}_b = \Lambda^{-\frac{1}{2}} Q^t S_b Q \Lambda^{-\frac{1}{2}}$ represents the between class matrix in the sphered space.

The above analysis shows that maximizing the Chernoff criterion produces linear discriminants corresponding to the largest values of $\bar{S}_b + V \Sigma V^t$ (Eq. (25)). It is straightforward to verify that if X indeed represents the solution to the Chernoff criterion (Eq. (9)), then $J_C(X) = \text{tr}(S_w^{-1} S_b) + \text{tr}(\tilde{\Lambda})$

The above shows that the Chernoff criterion is in a way very similar to the Fisher criterion. However, the Fisher criterion can not distinguish the difference between two discriminants whose generalized eigenvalues of $S_w^{-1} S_b$ are the same. On the other hand, the Chernoff criterion chooses the one whose difference between $\lambda^{(1)}$ and $\lambda^{(2)}$ is most pronounced. This can be seen from the fact that since

$$p_1 \lambda^{(1)} + p_2 \lambda^{(2)} = 1,$$

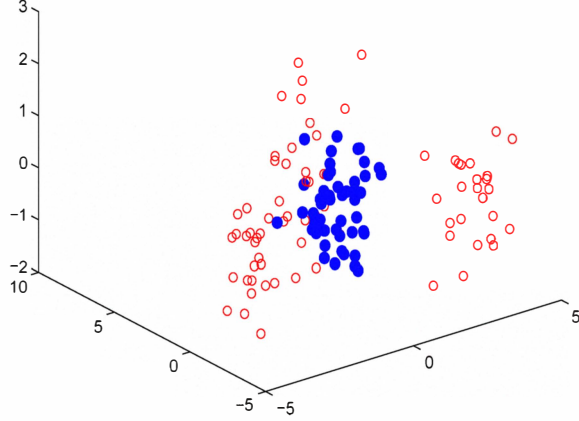


Fig. 2. A toy example, where the first class follows a Gaussian distribution with zero mean and covariance $0.5I$. The second class follows a mixture of three Gaussians, with means $[1 \ 4 \ 0]$, $[2\sqrt{3} \ -2 \ 0]$ and $[-2\sqrt{3} \ -2 \ 0]$, and covariance $0.5I$. The first class has 50 examples, while the second one has 75, with each mixture contributing 1/3 of total examples.

thus

$$\log\left(\frac{1}{(\lambda^{(1)})^{p_1}(\lambda^{(2)})^{p_2}}\right)$$

increases with either increasing $\lambda^{(1)}$ or $\lambda^{(2)}$. This is exactly what the Chernoff criterion is designed to do, e.g., to capture heteroscedasticity in data.

Another way to look at Eq. (25) is to examine the second term $\tilde{\lambda}$. Notice that $\tilde{\lambda}$ is derived from Eqs. (10) and (11). This process is known as FKT [26]. FKT has been shown to be an optimal reduced-rank representation under appropriate conditions [27]. While FKT relies entirely on exploiting the difference between within class matrices, Fisher's LDA mainly finds the discriminant subspace determined by the mean difference (after sphering data).

Our analysis shows that the Chernoff criterion is a combination of Fisher's LDA and FKT, thus capable of taking advantage of both worlds. When class means are identical (i.e., $S_b = 0$), the Fisher criterion can not be applied. On the other hand, the Chernoff criterion is applicable by replying on the difference in variance between classes. When within class matrices are equal (i.e., the same $\tilde{\lambda}_i$ values), FKT fails. In this case, the Chernoff criterion can find linear discriminants by exploring difference in class means (LDA).

In the case of $S_b \neq 0$, Chernoff can again differ from Fisher. Figure 1 illustrates a case in point. Since $S_b \neq 0$, Fisher chooses the direction along the Y axis as its discriminant. On the other hand, the difference between the two variances is the most along the X axis (i.e., more significant than the mean difference and the variance of the two classes along the Y axis combined). Thus, Chernoff selects the X axis as its discriminant.

It is also interesting to note that in the two class case, LDA can only obtain one dimensional projection, because $rank(S_b) = 1$ for two class problems. In contrast, the Chernoff criterion is capable of obtaining more than one discriminants, because the rank of Σ (Eq. (15)) is determined by the number of examples, not the number of classes.

The following simulated example (taken from [15]) illustrates a case in point. The left panel in Figure 3 shows a two class problem in three dimensions. The first class follows a Gaussian distribution with zero mean and covariance $0.5I$. The second class follows a mixture of three Gaussian distributions, with means $[1 \ 4 \ 0]$, $[2\sqrt{3} \ -2 \ 0]$ and $[-2\sqrt{3} \ -2 \ 0]$, and covariance $0.5I$. The first class has 50 examples, while the second one has 75, with each mixture contributing 1/3 of total examples. The left panel in Figure 3 shows the one dimensional projection obtained by LDA, where the two classes in the projected space overlap significantly. The right panel shows the two dimensional subspace computed by the Chernoff criterion, where there are two non-zero eigenvalues. This larger subspace provides a much better separation of the two classes.

V. ENHANCED VIEW OF DATA SPACE

Let A be the transformation such that $A^t(S_w + S_b)A = I$. It can be shown that $A^t S_w A$ and $A^t S_b A$ share the same eigen-space. That is,

$$A^t S_w A = B \Lambda_w B^t$$

$$A^t S_b A = B \Lambda_b B^t$$

and

$$\Lambda_w + \Lambda_b = I.$$

Furthermore, it is shown in [15] that if λ represents a generalized eigenvalue of $S_w^{-1}S_b$, then

$$\lambda = \frac{\lambda_b}{\lambda_w}. \quad (27)$$

In [15], the entire data space is decomposed into four xsubspaces based on Eq. (27), as shown in Figure 4. Here, the subspace 1 is the most discriminant, followed by subspaces 2 and 3. The common null space ($null(S_w) \cap null(S_b)$) does not contain any useful information.

We demonstrate a similar transformation to show that the above decomposition of the data space is incomplete. First, we let P be the transform such that

$$P^t(S_w + S_b + \tilde{S}_w)P = I.$$

Next, we simultaneously diagonalize $P^t S_w P$ and $P^t(S_b + \tilde{S}_w)P$

$$P^t S_w P = Z \Lambda_w Z^t$$

$$P^t(S_b + \tilde{S}_w)P = Z \Lambda_{bw} Z^t$$

and

$$\Lambda_w + \Lambda_{bw} = I.$$

As in [15], if λ represents an eigenvalue of $S_w^{-1}(S_b + \tilde{S}_w)$, then

$$\lambda = \frac{\lambda_{bw}}{\lambda_w}. \quad (28)$$

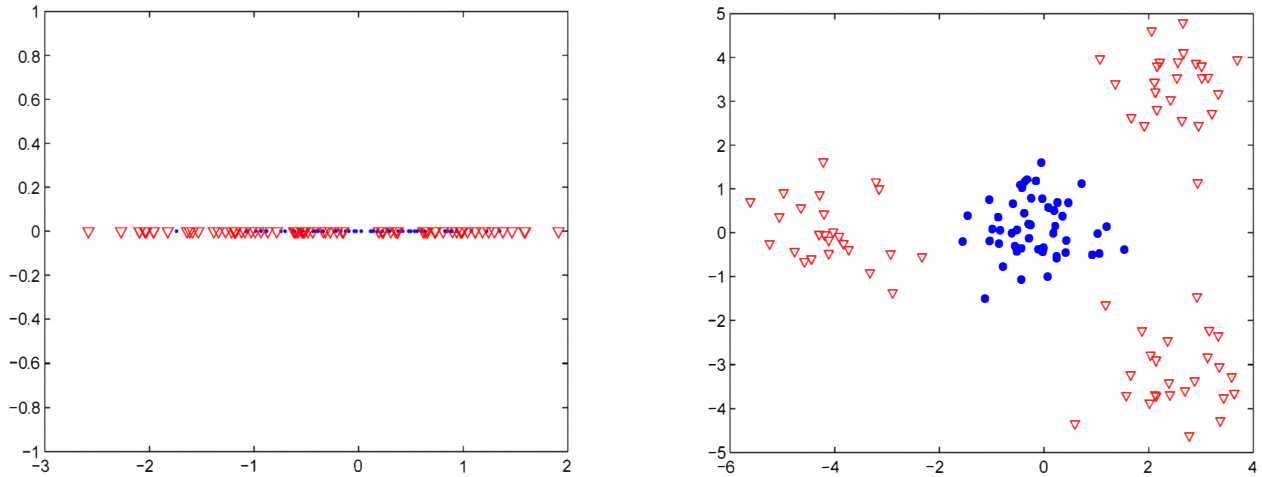


Fig. 3. Left panel: one dimensional projection obtained by LDA. Right panel: two dimensional subspace computed by the Chernoff criterion.

However, since both S_b and \tilde{S}_w contribute to λ_{bw} (in fact, S_b and \tilde{S}_w can be further diagonalized to show they share the same eigenspace), the decomposition of the entire data space into four subspaces suggested in [15] is incomplete according to the arguments we have presented here.

The characterization of discriminants residing in the data space should be much richer. The data space should be decomposed according to

$$\frac{\lambda_{bw}}{\lambda_w},$$

rather than

$$\frac{\lambda_b}{\lambda_w}.$$

Here we argue for at least an additional dimension that captures information provided by the difference in variance between classes (i.e., as represented by \tilde{S}_w Eq. (19)). This dimension tells us that given the same class mean difference, we prefer the discriminant along which within class matrices exhibit large variation. That is, this dimension should be measured by the eigenvalues of \tilde{S}_w (Eq. 19). This can be seen from Eqs. 21. It captures information about heteroscedasticity in the data, which is precisely what the Chernoff criterion is designed to do.

VI. EXPERIMENTS

Extensive experiments have been carried out comparing the Chernoff criterion against other competing methods [13], [22]. Since our purpose in this work is to provide theoretical insights into the Chernoff criterion, only a few experiments are performed here. Our experimental setup is very similar to that of [13].

We first apply PCA to training data to remove any principal components whose eigenvalue is smaller than one millionth of the total variance. This is to ensure that problems with near/or singular covariance matrices can be avoided and competing

transformations can be determined. In the transformed space, we use the one nearest neighbor classifier to determine accuracy.

We compare the following competing methods: Fisher criterion (Eq. (2)), Chernoff criterion (Eq. (9)), and FKT [3].

A. Data Sets

In these experiments, we compare Fisher, Chernoff, and FKT in two class classification problems. We use 9 data sets from the UCI machine learning database. They are all two class classification problems.

- **Glass Identification data (Glass)**. The data set has $n = 9$ continuous numerical features describing each of 214 instances in two classes: Window vs non-Window glasses. The objective is to assign the class label to each test instance. The average error rates are reported in the first row of Table I.
- **Wisconsin breast cancer data (Cancer Wisconsin)**. The data consists of 9 medical input features that are used to make a binary decision on the medical condition: determining whether the cancer is malignant or benign. The data set consists of 683 instances after removing missing values. The average error rates computed over all 2000 such classifications are reported in the second row in Table I.
- **Breast cancer data (Breast cancer)**. The data consists of 9 medical input attributes that are used to make a binary decision on the medical condition: determining whether the cancer is recurring (recurrence vs no-recurrence). The data set has 286 instances, of which 201 are in the no-recurrence class, while the remaining 85 are in the recurrence class. The average results are shown in the third row in Table I.
- **Heart disease diagnosis data (Heart Cleve)**. This data set consists of 303 instances in two classes (There are five original classes. However, we regrouped these five classes

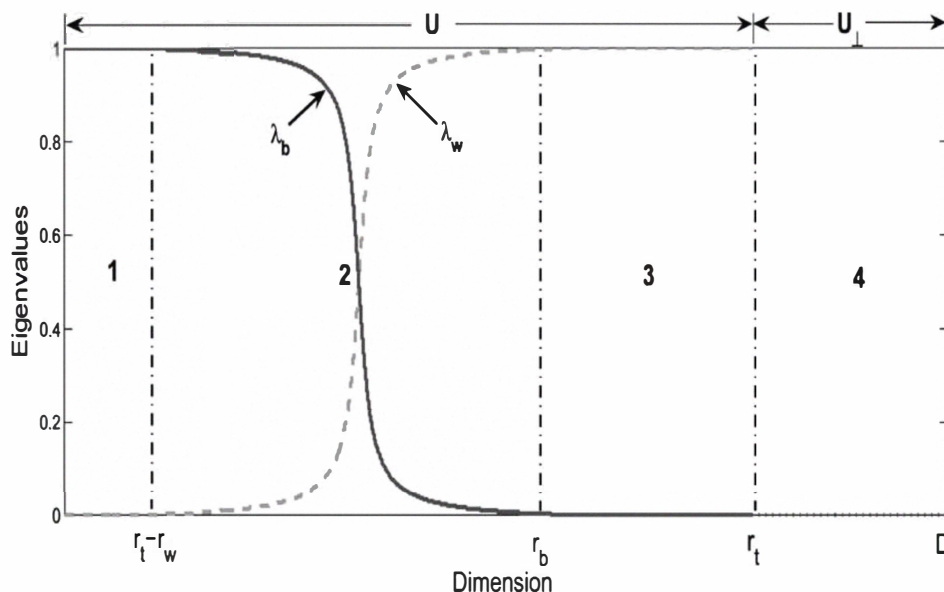


Fig. 4. The entire data space is decomposed into four subspaces via FKT as described in [15]. There is no discriminant information in U_{\perp} , the null space of $S_b + S_w$. In U , $\lambda_b + \lambda_w = 1$.

into two.) Each of these instances is represented by 13 numerical attributes. The data was collected at Cleveland Clinic Foundation. The goal is to predict the presence of heart disease in the patient. The average performance is shown in the fourth row in Table I.

- Heart disease diagnosis data (**Heart Hungary**). Similar to **Heart Cleve**, this data set consists of 294 instances represented by 13 numerical attributes. The data was collected at Hungarian Institute of Cardiology, Budapest. The objective is to predict whether a patient has heart disease. The average error rates are reported in the fifth row in Table I.
- Iris data (**Iris**). This data set consists of $n = 4$ measurements made on each of 100 iris plants of $J = 2$ species. The two species are iris versicolor and iris virginica. The problem is to classify each test point to its correct species based on the four measurements. The average error rates are shown in the sixth row of Table I.
- Letter data (**Letters**). This data set consists of a large number of black-and-white rectangular pixel arrays as one of the 26 upper-case letters in the English alphabet. Each letter is randomly distorted through a quadratic transformation to produce a set of 20,000 unique letter images that are then converted into $q = 16$ primitive numerical features. For this experiment we select letters “U” and “W”, where there are 813 “U” instances and 752 “W” instances. Thus, the data set consists of 1565 letter images.

- Pima Indians Diabete data (**Pima**). This data set consists of $n = 8$ numerical attributes measured for each of 768 samples of $J = 2$ classes. The problem is to classify each test point in the 8-dimensional space to its correct class. The average error rates over 20 independent runs are given in the eighth row in Table I.
- Ionosphere data (**Ionosphere**). The data consists of 34 electromagnetic features that are used to determine “good” or “bad” ($J = 2$) radar returns characterizing evidence of some type of structure in the ionosphere. The data set of 351 instances. Average error rates computed are reported in the last row in Table I.

For each data set, we randomly choose 60% as training and the remaining 40% as testing. We train Fisher, Chernoff and FKT on the training data and obtain projections. We then project both training and test data on the chosen subspace and use the 1-NN classifier to obtain average accuracy over ten runs. Note that for the two class case, one dimensional subspace is sufficient.

Table I shows the average accuracies registered by the three methods. Overall the Chernoff criterion generates good performance in the problems that we have experimented with. Our results are consistent with those provided in [13], [22].

VII. SUMMARY

This paper provides an analysis on the Chernoff criterion for linear dimensionality reduction. The Chernoff criterion has been proposed recently to address inability of LDA based on

TABLE I
CLASSIFICATION ERROR RATES IN SUBSPACES COMPUTED BY GDA AND
NDLA, USING 3-NN CLASSIFIER, ON 11 UCI DATA SETS.

Data Sets	Fisher	Chernoff	FKT
Glass	0.9024	0.9118	0.7094
Cancer Wisconsin	0.9538	0.9564	0.8249
Breast Cancer	0.6482	0.6618	0.6191
Heart Cleve	0.7703	0.7720	0.5763
Heart Hungary	0.7786	0.7547	0.6000
Iris	0.9400	0.9400	0.5625
Letters	0.9673	0.9679	0.6155
Pima	0.6827	0.6990	0.5772
Ionosphere	0.8193	0.8443	0.7571
average	0.8292	0.8342	0.6491

the Fisher criterion to deal with heteroscedasticity in data. The technique extends well-known Fisher's LDA and is capable of exploiting heteroscedasticity in data. While the Chernoff criterion is shown to outperform the Fisher criterion, a clear understanding of its exact behavior is lacking. In addition, the criterion, as introduced, is rather complex, making it difficult to clearly state its relationship to other linear dimensionality reduction techniques. In this paper, we have shown precisely what can be expected from the Chernoff criterion and its relations to the Fisher criterion and FKT. In addition, we have shown that a recently proposed decomposition of the data space into four subspaces is incomplete. We have provided evidence on how to best enrich the decomposition of the data space to account for heteroscedasticity in the data.

In this paper, our focus is on the Chernoff criterion for the two class case. While our analysis for the two class case provides a clue on its behavior in multiclass problems, a direct analysis is highly desirable, which we intend to pursue in our future work.

REFERENCES

- [1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.
- [2] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag: New York, 1986.
- [3] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990.
- [4] V. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [5] T. Cootes and C. Taylor, "Active shape models: Smart snakes," in *Proc. British Machine Vision Conf.*, 1992, pp. 9–18.
- [6] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human faces," in *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 2148–2151.
- [7] P. Howland and H. Park, "Generalizing discriminant analysis using the generalized singular value decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 995–1006, 2004.
- [8] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.
- [9] S. Nayar, S. Baker, and H. Murase, "Parametric feature detection," in *IEEE Conf. Computer Vision and Pattern Recognition*, 1994, pp. 471–477.
- [10] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 84–91.
- [11] M. Turk and A. Pentland, "Eigenfaces for recognition," *Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [12] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems," *Journal of Machine Learning Research*, vol. 6, pp. 483–502, 2005.
- [13] M. Loog and P. Duin, "Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 732–739, 2004.
- [14] M. Loog, *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion*. Delft Univ. Press, 1999.
- [15] S. Zhang and T. Sim, "Discriminant subspace analysis: A fukunaga-koontz approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1732–1745, 2007.
- [16] H. Decell and S. Mayekar, "Feature combinations and the divergence criterion," *Computers and Math. with Applications*, vol. 3, pp. 71–76, 1977.
- [17] P. Devijver and J. K. Pattern, *Recognition: A Statistical Approach*. London: Prentice-Hall, 1982.
- [18] N. Kumar and A. Andreou, "Generalization of linear discriminant analysis in a maximum likelihood framework," in *Proceedings of Joint Meeting of the Am. Statistical Assoc.*, 1996.
- [19] Y. Sun and D. Wu, "A relief based feature extraction algorithm," in *Proceedings of SIAM International Conference on Data Mining*, 2008, pp. 188–195.
- [20] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin based feature selection - theory and algorithms," in *Proceedings of 21st International Conference on Machine Learning*, 2004, pp. 43–50.
- [21] W. Zhang, X. Xue, Z. Sun, Y. Guo, and H. Lu, "Optimal dimensionality of metric space for classification," in *Proceedings of 24th International Conference on Machine Learning*, 2007, pp. 1135–1142.
- [22] L. Rueda and M. Herrera, "Linear dimensionality reduction by maximizing the chernoff distance in the transformed space," *Pattern Recognition*, vol. 41, no. 10, pp. 3138–3152, 2008.
- [23] O. Hamsici and A. Martinez, "Bayes optimality in linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 647–657, 2008.
- [24] C. Chen, "On information and distance measures, error bounds, and feature selection," *The Information Scientist*, vol. 10, pp. 159–173, 1979.
- [25] J. Chung, P. Kannappan, C. Ng, and P. Sahoo, "Measures of distance between probability distributions," *J. Math. Analysis and Applications*, vol. 138, pp. 280–292, 1989.
- [26] F. Fukunaga and W. Koontz, "Applications of the karhunen-loeve expansion to feature selection and ordering," *IEEE Transactions on Computers*, vol. 19, no. 5, pp. 311–318, 1970.
- [27] X. Huo and et al, "Optimal reduced-rank quadratic classifiers using the fukunaga-koontz transform, with applications to automated target recognition," in *Proc. of SPIE Conference*, 2003.