

Chernoff Dimensionality Reduction—Where Fisher Meets FKT *

Jing Peng[†] Guna Seetharaman[‡] Wei Fan[§] Stefan Robila[¶] Aparna Varde^{||}

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, making it difficult to clearly state its relationship to other linear dimensionality reduction 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. Finally, we provide experimental results validating our theoretical analysis.¹ **Keywords:** Dimension reduction, LDA, FKT, Chernoff distance.

1 Introduction

In classification, a large number of features or attributes often make the design of a classifier difficult and degrades its performance². 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, subspace methods play an important role by significantly reducing the number of features for building classifiers. For example, in visual learning and modeling the principal modes are extracted and utilized for description, detection, and classification. Using these principal modes to represent data can be found in parametric descriptions of shape [2], object detection [3], visual learning [4], face recognition [5, 6], and Fisherfaces [7].

There are many dimensionality reduction techniques for classification in the literature. The two popular ones are Fisher’s linear discriminant analysis (LDA) [8] and Fukunaga-Koontz Transform (FKT) [9]. FKT has shown promise in vision and classification applications [10, 11]. It can be shown that under appropriate conditions FKT is an optimal reduced-rank representation [10]. Furthermore, FKT does not suffer from the small sample size problem often associated with LDA. FKT assumes that target and clutter objects have the same mean. Therefore, it relies entirely on the difference in variance between target and clutter to compute reduced-rank representations. However, it may not be adequate in many applications.

LDA, on the other hand, simply tries to separate class means as much as possible. In LDA, we are given a set of l examples: $z = \{(x_i, y_i)\}_{i=1}^l$. 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, $x_i \in X \subset \mathfrak{R}^q$ are the q -dimensional inputs, and $y_i \in Y = [-M, M] \subset \mathfrak{R}$ are scalar labels. According to Fisher’s criterion, one has to find a projection matrix $W \in \mathfrak{R}^{q \times d}$ that maximizes:

$$(1.1) \quad J_F(W) = \text{tr}(W^t S_w W)^{-1} W^t S_b W$$

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 the so-called between-class and within-class matrices. Here m represents the overall mean, m_c denotes the mean of class

*^(a) A part of the research described in this material is based upon work funded by AFRL, under AFRL Contract No. FA8750-09-2-0155. (b) Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFRL.

[†]Computer Science Department, Montclair State University, Montclair, NJ 07043; jing.peng@montclair.edu.

[‡]Information Directorate, AFMC AFRL/RITB, Rome, NY; Gunasekaran.Seetharaman@rl.af.mil.

[§]IBM T.J. Watson Research; weifan@us.ibm.com.

[¶]Computer Science Department, Montclair State University, Montclair, NJ 07043; stefan.robila@montclair.edu.

^{||}Computer Science Department, Montclair State University, Montclair, NJ 07043; aparna.varde@montclair.edu.

¹Part of Abstract appeared in [14].

²Part of Introduction appeared in [14].

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 the 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 [9].

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 [12]. The technique makes use of directed distance matrices (DDMs) [13], 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 reduction 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 and the Fukunaga-Koontz transform (FKT) [9]. In fact, we show that in the two class case, when two classes have two different means, the Chernoff criterion demonstrates the characteristics of both Fisher and FKT. On the other hand, when the two classes have the same mean, the Chernoff criterion reduces to FKT. Thus, the Chernoff criterion takes advantage of both worlds. Furthermore, we show that the decomposition of the data space into four subspaces described in [11] is incomplete. We provide arguments on how to enrich the decomposition of the data space to account for heteroscedasticity in the data. In this work, we focus on two class problems. This can be justified by the fact that Chernoff distance is intended for two distributions. Finally, we provide experimental results validating our theoretical analysis.

We state at the outset that this work has significantly extended our earlier analysis of Chernoff dimensionality reduction that appeared in [14] in the following ways: (1) This work provides a significantly precise statement on the interplay among Chernoff, Fisher and FKT through complete mathematical analysis (Sections 4.1, 4.2, and 4.3), which is not the case in our earlier work [14]; (2) This work provides a significantly clear augmentation to data space decomposition [11], in particular how subspace 3 should be augmented (Section 6), which is not the case in our earlier work [14]; and (3) This work provides examples that demonstrate clearly how Chernoff criterion is related to Fisher and FKT (Figures 1 and 2, and Section 5), which is not the case in our earlier work [14].

2 Related Work

Note that part of Related Work appeared in [14]. Several methods for extending linear classifiers to unequal covariance matrices and non Gaussian distributions are discussed in [9]. 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-Leibler divergence to extend the Fisher criterion is proposed in [15]. Several techniques based on probabilistic separability and interclass distance measures can be introduced in [16]. These techniques require rather time-consuming iterative procedures to compute linear discriminants.

A maximum-likelihood approach to LDA is described in [17]. 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.

A dimensionality reduction technique that improves the Fisher criterion in heteroscedastic data has been proposed [12]. 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 [18]. 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 [19].

A large number of subspace methods have been proposed, most of which address the computational

difficulty associated with LDA when the small sample size problem occurs (S_w becomes singular). PCA+LDA uses the pseudo-inverse S_w^+ in place of S_w^{-1} . However, computing S_w^+ is ill posed. Another method is to use PCA to remove the null space of S_w , and then apply LDA to the reduced representation. However, this method remains sub-optimal because the null space of S_w can contain discriminant information [20].

newLDA [20] first transforms the data into the null space of S_w . It then applies PCA to maximize the between-class matrix in the transformed space. Its performance degrades with decreasing dimensions of the null space. A variant of LDA+PCA is proposed in [21]. The method first discards the null space of $S_w + S_b$ that is the common null space of both S_w and S_b . And as such, discarding this null space does not lose any discriminant information. The method then applies LDA+PCA to the reduced representation in the transformed space.

Weighted piecewise LDA is another technique for addressing the small sample size problem [22]. The technique first creates subsets of features and applies LDA to each subset. The technique then combines the resulting piecewise linear discriminants to produce an overall solution. More recently, discriminant analysis based on the average margin is proposed [23]. The technique is closely related to LDA but does not involve inverting matrices. Since the criterion ($tr(S_b - S_w)$) is additive, the technique does not suffer from the small sample size problem.

In [24], a two-stage LDA technique is proposed. This technique not only avoids the small sample size problem of LDA but also achieves greater computational efficiency. This is accomplished by applying QR decomposition first, followed by LDA. In [25], the small sample size problem is addressed by simultaneously diagonalizing the between- and within-class scatter matrices through generalized singular value decomposition (GSVD) [26, 27]. This technique also achieves greater computational efficiency. In [11], a clear connection between GSVD and FKT for the LDA problem has been established.

A dimension reduction technique, called linear feature extraction (LFE), based on Relief [28] is introduced [29]. In [30], a metric space dimension reduction technique is proposed. 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.

A dimension reduction technique based on max-min distance analysis has been proposed recently [31]. For a multi-class problem with homoscedastic Gaussian distributions, this technique computes discriminants by

maximizing the minimum pairwise distance between classes. The idea is that maximizing the minimum pairwise distance produces a subspace that is overall more discriminant. This is similar to the argument made in support vector machines, where maximizing the minimum margin results in better generalization [32]. The experimental results presented in [31] show that max-min distance analysis is promising.

In [33], geometric means for subspace selection are investigated. It is shown that in a multi-class problem, when all covariances are Gaussian and identical, the Fisher criterion is equivalent to the one that maximizes the KL divergence of the classes. Several criteria based on the geometric mean of KL divergence are studied and empirically compared against competing techniques, including the Chernoff criterion. It turns out the proposed geometric mean criteria are very competitive in the problems experimented.

3 Chernoff Criterion

The material presented in this section is taken from our earlier work that appeared in [14].

The Fisher criterion (Eq. (1.1)) 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 $S_b w = \lambda S_w w$ [9].

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 :

$$(3.2) \quad D_C = -\log \int p_1^\alpha(x) p_2^{1-\alpha}(x) dx$$

where $\alpha \in \{0, 1\}$ is a constant. For two normally distributed densities, the Chernoff distance can be written as [34, 35]

$$D_C = (m_1 - m_2)^t (p_1 S_1 + p_2 S_2)^{-1} (m_1 - m_2)$$

$$(3.3) \quad + \frac{1}{p_1 p_2} \log \frac{|p_1 S_1 + p_2 S_2|}{|S_1|^{p_1} |S_2|^{p_2}}.$$

where $S_i = \sum_{x_j, y_j=i} (x_j - m_i)(x_j - m_i)^t$ for $i = 1, 2$. It can be shown that one can obtain D_C as the trace of matrix S_C [12]

$$S_C = S^{-1/2} S_e S^{-1/2} + \frac{1}{\alpha\beta} (\log S - \alpha \log S_1 - \beta \log S_2),$$

where $\beta = 1 - \alpha$. 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:

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

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

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

We will then transform this back to the original space by $S_w^{1/2}$. For the Fisher criterion, we have [12]

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

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

The heteroscedastic two-class Chernoff criterion J_C is defined as

$$(3.4) \quad \begin{aligned} J_C(W) = & \text{tr}((W^t S_w W)^{-1} (W^t S_b W \\ & - W^t S_w^{\frac{1}{2}} (p_1 \log(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}) \\ & + p_2 \log(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}})) S_w^{\frac{1}{2}} W)) \end{aligned}$$

4 Chernoff Dimension Reduction: Combining Fisher and Fukunaga-Koontz Transform

We note that part of this section appeared in our earlier work [14]. However, subsections 4.1, 4.2, and 4.3 are not. These sections provide analysis that significantly extended our earlier work in terms of precise characterization of the interplay among Chernoff, Fisher and FKT.

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) [36]. Since $S_w = p_1 S_1 + p_2 S_2$, we can simultaneously diagonalize S_1 and S_2 . Let

$$(4.5) \quad P = Q\Lambda^{-\frac{1}{2}}$$

where
(4.6)

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

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 [9]

$$(4.7) \quad \tilde{S}_1 = V\Lambda^{(1)}V^t$$

and

$$(4.8) \quad \tilde{S}_2 = V\Lambda^{(2)}V^t$$

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

$$(4.9) \quad \begin{aligned} S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}} &= Q P^t S_1 P Q^t \\ &= Q \tilde{S}_1 Q^t \\ &= Q V \Lambda^{(1)} V^t Q^t \end{aligned}$$

Since V and Q are orthogonal, it follows that

$$(4.10) \quad \begin{aligned} p_1 \log(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}) &= p_1 \log(Q V \Lambda^{(1)} S_1 V^t Q^t) \\ &= p_1 Q V \log(\Lambda^{(1)}) V^t Q^t \\ &= p_1 Q V \tilde{\Lambda}^{(1)} V^t Q^t, \end{aligned}$$

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

$$(4.11) \quad p_2 \log(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}}) = p_2 Q V \tilde{\Lambda}^{(2)} V^t Q^t,$$

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

$$(4.12) \quad \Sigma = p_1 \log(S_w^{-\frac{1}{2}} S_1 S_w^{-\frac{1}{2}}) + p_2 \log(S_w^{-\frac{1}{2}} S_2 S_w^{-\frac{1}{2}})$$

Combining Eqs. (4.10) and (4.11) gives rise to

$$(4.13) \quad \begin{aligned} \Sigma &= p_1 Q V \tilde{\Lambda}^{(1)} V^t Q^t + p_2 Q V \tilde{\Lambda}^{(2)} V^t Q^t \\ &= Q V (p_1 \tilde{\Lambda}^{(1)} + p_2 \tilde{\Lambda}^{(2)}) V^t Q^t \\ &= Q V \tilde{\Lambda} V^t Q^t, \end{aligned}$$

where

$$\begin{aligned} \tilde{\Lambda}_i &= p_1 \bar{\Lambda}_i^{(1)} + p_2 \bar{\Lambda}_i^{(2)} \\ (4.14) \quad &= \log((\lambda_i^{(1)})^{p_1} (\lambda_i^{(2)})^{p_2}) \end{aligned}$$

Define

$$(4.15) \quad \tilde{S}_w = S_w^{\frac{1}{2}} \Sigma S_w^{\frac{1}{2}}.$$

Then the Chernoff criterion becomes

$$(4.16) \quad J_C(W) = \text{tr}((W^t S_w W)^{-1} (W^t (S_b - \tilde{S}_w) W)).$$

We can optimize $J_C(W)$ by solving an eigenvalue decomposition of the matrix

$$(4.17) \quad S_w^{-1} (S_b - \tilde{S}_w).$$

In the following sections, we analyze how each of the terms in (4.17) contributes to Chernoff dimensionality reduction.

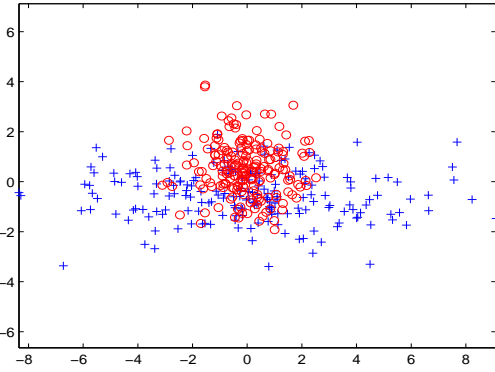


Figure 1: In this two class example, the difference between the two co-variances is most pronounced along the X axis, while the variances along the Y axis are the same. Fisher chooses the Y axis as its discriminant, while Chernoff chooses the X axis.

4.1 Chernoff and FKT We begin by examining the second term in (4.17) and the role it plays in computing Chernoff discriminants. We do so by analyzing the solution to Eq. (4.17), and thus Eq. (3.4). First, we rewrite $\tilde{\Lambda}_i$ in Eq. (4.14) as

$$(4.18) \quad \tilde{\lambda}_i = \tilde{\Lambda}_i = \log\left(\frac{1}{(\lambda_i^{(1)})^{p_1} (\lambda_i^{(2)})^{p_2}}\right).$$

Eq. (4.17) becomes

$$(4.19) \quad S_w^{-1} (S_b + \tilde{S}_w)$$

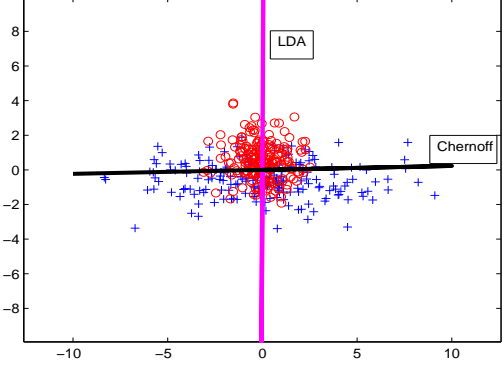


Figure 2: When both Fisher and Chernoff were applied to the example in Fig 1, Fisher chose the Y axis as its discriminant, while Chernoff chose the X axis.

Thus, there are two terms contributing to Chernoff criterion. The first term, $S_w^{-1} S_b$, is the classic Fisher's criterion, while the second term, $S_w^{-1} \tilde{S}_w$, requires further discussion. From (4.6), (4.13), and (4.15), we have

$$(4.20) \quad S_w^{-1} \tilde{S}_w = S_w^{-1/2} \tilde{S}_w S_w^{1/2}$$

$$(4.21) \quad = Q \Lambda^{-1/2} V \tilde{\Lambda} V^t \Lambda^{1/2} Q^t$$

$$(4.22) \quad = Z^{-1} \tilde{\Lambda} Z$$

where $Z = V^t \Lambda^{1/2} Q^t$. Thus, $S_w^{-1} \tilde{S}_w$ and $\tilde{\Lambda}$ are similar under similarity transformation Z . Furthermore, $S_w^{-1} \tilde{S}_w$ and $\tilde{\Lambda}$ share the same eigenvalues.

When class means are identical (i.e., $S_b = 0$), Fisher's criterion can not be applied. $S_w^{-1} (S_b + \tilde{S}_w)$ reduces to

$$S_w^{-1} \tilde{S}_w = Z^{-1} \tilde{\Lambda} Z.$$

Thus, the Chernoff criterion selects the eigenvector having the largest eigenvalue of $\tilde{\Lambda}$ as its discriminant. Since $p_1 \lambda^{(1)} + p_2 \lambda^{(2)} = 1$, or equivalently $\lambda^1 + \lambda^2 = 1$, where $\lambda^1 = p_1 \lambda^{(1)}$ and $\lambda^2 = p_2 \lambda^{(2)}$ and

$$(4.23) \quad \tilde{\lambda} = \log\left(\frac{1}{(\lambda^{(1)})^{p_1} (\lambda^{(2)})^{p_2}}\right)$$

$$(4.24) \quad = \log\left(\frac{1}{(\lambda^1)^{p_1} (1 - \lambda^1)^{p_2}}\right),$$

the largest $\tilde{\lambda}$ corresponds to the largest difference between $\lambda^{(1)}$ and $\lambda^{(2)}$. Thus, $\tilde{\lambda}$ increases with increasing $\lambda^{(1)}$ (decreasing $\lambda^{(2)}$) or decreasing $\lambda^{(1)}$ (increasing $\lambda^{(2)}$). This is exactly what the Chernoff criterion is designed to do, e.g., to capture heteroscedasticity in data.

This behavior of Chernoff dimension reduction is closely related to FKT [9]. For the purpose of our discussion, we use 1 and 2 to denote target class +1

and clutter class -1, respectively. For the moment, we assume both target and clutter follow Gaussian distributions:

$$P_1(x) = \propto \exp\{-(x - m_1)^t S_1^{-1}(x - m_1)\}$$

and

$$P_2(x) = \propto \exp\{-(x - m_2)^t S_2^{-1}(x - m_2)\}.$$

The optimal maximum likelihood classifier is given by

$$(4.25) \quad f(x) = \frac{P_1(x)}{P_2(x)},$$

which is proportional to

$$\exp\{-(x - m_1)^t S_1^{-1}(x - m_1) + (x - m_2)^t S_2^{-1}(x - m_2)\}.$$

Thus, x is classified as target if $f(x)$ is greater than a threshold, and clutter otherwise.

Since the sum of the matrices $S_1 + S_2$ is positive semi-definite and thus can be factorized into

$$S_1 + S_2 = \Phi D \Phi^t.$$

Letting $P = \Phi D^{-1/2}$, we have

$$(4.26) \quad P^t(S_1 + S_2)P = I.$$

Now let $T = P^t S_1 P$ and $C = P^t S_2 P$. It follows that $T + C = I$. And define $G_1 = P^t(x - m_1)$ and $G_2 = P^t(x - m_2)$. We obtain

$$(4.27) \quad f(x) = c \exp\{-G_1^t T^{-1} G_1 + G_2^t C^{-1} G_2\},$$

where c is constant. Now suppose $m_1 = m_2 = m$ (assumption of FKT). We then have $G_1 = G_2 = G$. Thus, the optimal classifier thus becomes

$$\text{Target} = \mathbf{1}(c \exp\{-G^t(T^{-1} - C^{-1})G\} \geq \alpha)$$

where $\mathbf{1}(\cdot)$ is the indicator (i.e., $\mathbf{1}(\cdot)$ is 1 if its argument is true, and 0 otherwise), and α is a constant threshold.

Since T and C share the same eigen space,

$$T = \Theta^t \Lambda \Theta \quad \text{and} \quad C = \Theta^t (I - \Lambda) \Theta,$$

we have

$$f(x) = C \exp\{-(\Theta^{-t} G)^t (\Lambda^{-1} - (I - \Lambda)^{-1}) (\Theta^{-t} G)\}.$$

Define $W = \Theta^{-t} G$. $f(x)$ can be further simplified as

$$f(x) = C \exp\{-W^t (\Lambda^{-1} - (I - \Lambda)^{-1}) W\}.$$

Or equivalently, we can write

$$(4.28) \quad \begin{aligned} g(x) &= -W^t (\Lambda^{-1} - (I - \Lambda)^{-1}) W \\ &= \sum_{i=1}^d \left(\frac{1}{1 - \lambda_i} - \frac{1}{\lambda_i} \right) w_i^2. \end{aligned}$$

This is FKT. Classification can be made according to

$$\text{Target} = \mathbf{1}\left(\sum_{i=1}^d \left(\frac{1}{1 - \lambda_i} - \frac{1}{\lambda_i}\right) w_i^2 \geq \alpha\right).$$

Notice that all dimensions are used in the decision function in Eq. (4.28). In order to achieve a reduced rank representation of target and clutter objects, tuned basis functions (TBFs) choose the dimensions (eigenvectors) having $\max|\lambda - 1/2|$, resulting in a reduced rank representation. The dimensions having $\lambda > 1/2$ represent target features, while those with $\lambda < 1/2$ represent clutter features. This exactly mirrors the behavior of $\tilde{\lambda}$ in (4.24). Thus, when $S_b = 0$, the second term in (4.19) has been shown to be an optimal reduced-rank representation under appropriate conditions [10].

4.2 Chernoff and Fisher Suppose

$$(4.29) \quad \lambda_i^{(1)} = \lambda_i^{(2)} = \lambda_i$$

for all i s. We have

$$(4.30) \quad \begin{aligned} 1 &= p_1 \lambda_i^{(1)} + p_2 \lambda_i^{(2)} \\ &= (p_1 + p_2) \lambda_i \\ &= \lambda_i. \end{aligned}$$

Here we used the fact that $p_1 + p_2 = 1$. Thus, from (4.14)

$$(4.31) \quad \begin{aligned} \tilde{\Lambda}_i &= \log((\lambda_i^{(1)})^{p_1} (\lambda_i^{(2)})^{p_2}) \\ &= \log((\lambda_i)^{p_1} (\lambda_i)^{p_2}) \\ &= \log((1)^{p_1} (1)^{p_2}) \\ &= 0. \end{aligned}$$

It follows that

$$(4.32) \quad \begin{aligned} \tilde{S}_w &= Q \Lambda^{1/2} V \tilde{\Lambda} V^t \Lambda^{1/2} Q^t \\ &= 0. \end{aligned}$$

The Chernoff criterion thus becomes

$$(4.33) \quad S_w^{-1}(S_b - \tilde{S}_w) = S_w^{-1} S_b.$$

That is, when covariances are the same for two classes, the Chernoff criterion reduces to the Fisher criterion, as expected.

4.3 Chernoff: Combining Fisher and FKT 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

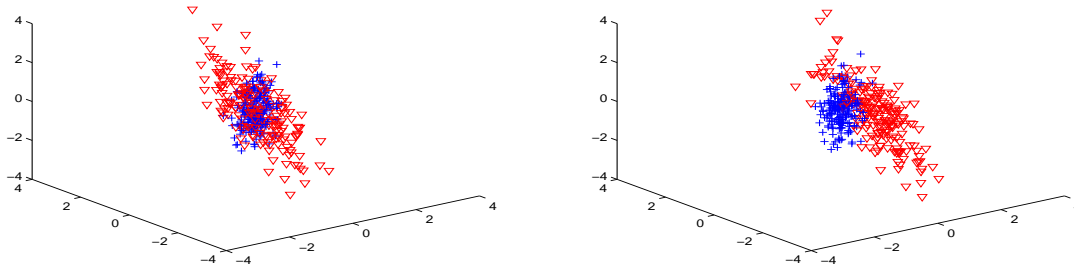


Figure 3: Two simple simulated examples in three dimensions. Left: Two Gaussians share the same zero mean but different covariance matrices. Right: Two Gaussians have different means (0,0,0) and (1,0,-0.5).

by replying on the difference in variance between classes. When within class matrices are equal (i.e., $\lambda_i^{(1)} = \lambda_i^{(2)}$), FKT fails. In this case, the Chernoff criterion reduces to Fisher’s criterion and can find linear discriminants by exploring difference in class means (LDA).

When neither $S_b = 0$ nor $\lambda_i^{(1)} = \lambda_i^{(2)}$, Chernoff can again differ from Fisher (it will certainly be different from FKT). Figure 1 illustrates a case in point. In this two class example, the two class means are different only along the Y axis. Also, as can be seen, the difference between the two co-variances is most pronounced along the X axis, while the variances along the Y axis are the same. Fisher chooses the Y axis as its discriminant, while Chernoff chooses the X axis. When both Fisher and Chernoff were applied to the example, Fisher chose the direction along the Y axis as its discriminant, since $S_b \neq 0$. 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 chose the X axis as its discriminant, as expected.

5 Simple Illustration

To gain an intuitive understanding of the Chernoff criterion and its relations to Fisher and FKT, we begin with two simple simulated examples.

- **Mixed:** Two Gaussian classes: same mean but different covariance matrices. The two classes share the same zero mean in three dimensional space. Each class has 200 samples. The two covariance matrices are

$$C_1 = [1, 1, 0]' * [1, 1, 0] + 0.1 * [0, 1, 1]' * [0, 1, 1]$$

and

$$C_2 = [0, 1, 1]' * [0, 1, 1] + 0.1 * [1, 0, 1]' * [1, 0, 1].$$

- **Separated:** This example is the same as the previous one, except the two classes have different means: (0,0,0) and (1,0,-0.5).

Figure 3 shows the two examples. Clearly, the Fisher criterion will fail in the first example, since $S_b = 0$. However, the actual means estimated from the samples may be different. The first example clearly favors techniques that exploit differences in variance such as FKT, while the second example favors techniques that rely on mean differences such as LDA.

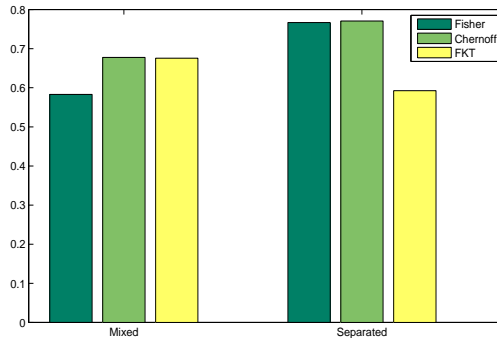


Figure 4: Average accuracy in the subspace computed by the competing methods using the one nearest neighbor rule, on two simulated data sets.

Fig. 4 shows the average accuracy registered by the three methods on the two exmples. As expected, Fisher did poorly on **Mixed**, while FKT shows poor performance on the **Separated** example. On the other hand, Chernoff encompasses the strengths of both Fisher and FKT. It therefore performs well on both examples.

Figure 5 shows one dimensional data projection by the three competing methods on the two simulated

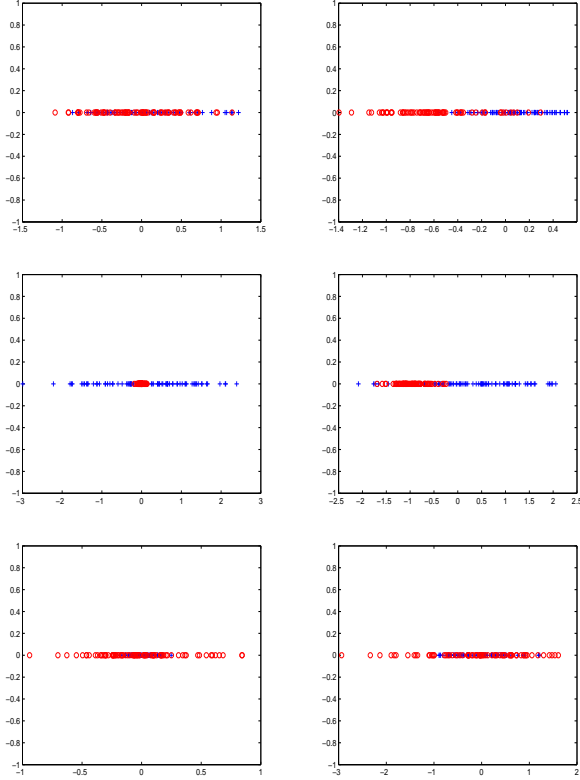


Figure 5: One dimensional data projection by the three methods on the two simulated problems (Mixed vs Separated). Top: Fisher criterion. Middle: Chernoff criterion. Bottom: FKT.

problems. Top row: One dimensional projections by Fisher on the **Mixed** example (left) and the **Separated** example (right). Similarly, The middle and bottom row show the projections by Chernoff and FKT on the two examples. These results again demonstrate the advantage of the Chernoff method.

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 $S_w^{-1}(S_b + \tilde{S}_w)$ (Eq. (4.19)) is determined not only by the number of classes, but also by differences in variance between the two classes.

The following simulated example (taken from [11]) illustrates a case in point. While this example also appeared in [14], the comparison with FKT was not provided. Figure 6 shows a two class problem in three dimensions. The first class follows a Gaussian distribution with zero mean and covariance $0.5\mathbf{I}$. The second class follows a mixture of three Gaussian distributions,

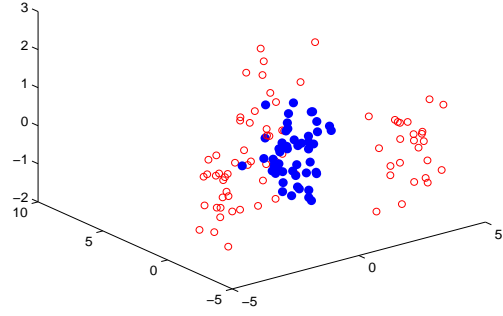


Figure 6: A three dimensional toy example, where the first class follows a Gaussian distribution with zero mean and covariance $0.5\mathbf{I}$. The second class follows a mixture of three Gaussians, with means $[1\ 4\ 0]$, $[2\sqrt{3}\ 0]$ and $[-2\sqrt{3}\ 0]$, and covariance $0.5\mathbf{I}$. The first class has 50 examples, while the second one has 75, with each mixture contributing 1/3 of total examples.

with means $[1\ 4\ 0]$, $[2\sqrt{3}\ 0]$ and $[-2\sqrt{3}\ 0]$, and covariance $0.5\mathbf{I}$. The first class has 50 examples, while the second one has 75, with each mixture contributing 1/3 of total examples.

The top left panel in Figure 7 shows the one dimensional projection obtained by LDA, where the two classes in the projected space overlap significantly. The top right panel shows the dimensional projection computed by FKT. In this example, FKT provides superior one dimensional projection over LDA. The bottom right panel shows 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. The bottom left panel shows a two dimensional projection by FKT. Chernoff provides superior separation over FT.

6 Enhanced View of Data Space

We note that the first paragraph in this section is taken from our earlier work appeared in [14]. 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 eigenspace. That is, $A^t S_w A = B\Lambda_w B^t$ and $A^t S_b A = B\Lambda_b B^t$. Furthermore,

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

It is shown in [11] that if λ represents a generalized eigenvalue of $S_w^{-1}S_b$, then

$$(6.34) \quad \lambda = \frac{\lambda_b}{\lambda_w},$$

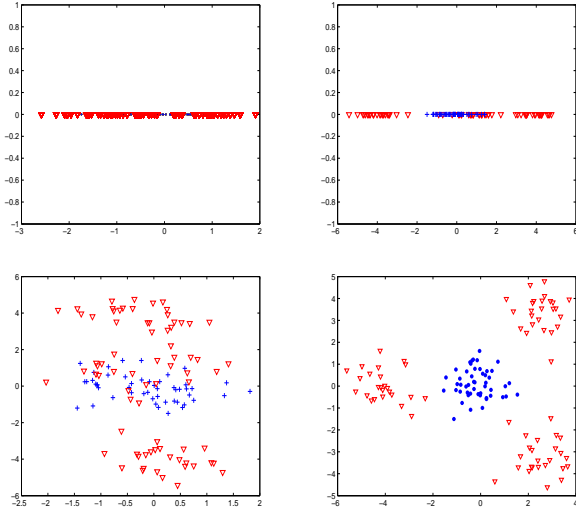


Figure 7: Top panel: one dimensional projections obtained by LDA (left) and FKT (right). Bottom panel: two dimensional subspaces computed by FKT (left) and the Chernoff criterion (right).

where λ_w is an eigenvalue of S_w , and λ_b is an eigenvalue of S_b .

In [11], the entire data space is decomposed into four subspaces based on Eq. (6.34), as shown in Figure 8. Here, subspace 1 is the most discriminant, followed by subspaces 2 and 3. The common null space ($null(S_w) \cap null(S_b)$), subspace 4, does not contain any useful information.

Our analysis shows that the above decomposition of the data space is incomplete. Clearly, subspace 4 contains no useful information, and thus can be safely discarded. Also, the analysis of subspace 1 is quite clear. Given that $\lambda_w = 0$, any thresholding along $m_1 - m_2$ suffices. Thus, no further analysis is necessary. On the other hand, subspace 2 is complex, where neither $\lambda_b = 0$ nor $\lambda_w = 0$. Let $\lambda_l^{Chernoff}$ be the largest eigenvalue of $S_w^{-1}(S_b + \tilde{S}_w)$, λ_l^{Fisher} be the largest eigenvalue of $S_w^{-1}S_b$, and $\tilde{\lambda}_l$ be the largest eigenvalue of $S_w^{-1}\tilde{S}_w$. Then

$$\lambda_l^{Chernoff} \leq \lambda_l^{Fisher} + \tilde{\lambda}_l.$$

Thus, for a fixed λ_b or λ_w , the dependence of $\lambda_l^{Chernoff}$ on λ_l^{Fisher} and $\tilde{\lambda}_l$ will be application specific. It is difficult to obtain a general statement about subspace 2. Therefore, in this work, we focus on subspace 3.

Let us consider subspace 3, where $\lambda_b = 0$. Subspace 3 corresponds to the case where class means are identical, i.e., $S_b = 0$. In this case, neither LDA nor LDA/FKT [11] is capable of finding a solution. On the other hand, the Chernoff criterion reduces to FKT when

$S_b = 0$. Thus, Chernoff simply computes discriminants by exploiting the difference in variance between the two classes, as in FKT. This shows that subspace 3 contains more information than what is shown in Fig. 8.

Our analysis indicates that not every dimension in subspace 3 is equally discriminant. The characterization of discriminants residing in subspace 3 should be much richer. Thus, according to Chernoff (4.19), subspace 3 should be augmented by $\lambda^{(1)}$ (or $\lambda^{(2)}$) that indicates the usefulness of a dimension in the space, as shown in Figure 9. Here, any dimension in subspace 3 whose corresponding $\tilde{\lambda}$ is large is more discriminant.

The above augmentation tells us that given the same class means, we prefer the discriminant along which within-class matrices exhibit the largest variation. That is, this dimension should be measured by the eigenvalues of \tilde{S}_w (Eq. 4.15). It captures information about heteroscedasticity in the data, which is precisely what the Chernoff criterion is designed to do.

7 Experiments

Notice that the following standard data description also appeared in [14].

Extensive experiments have been carried out comparing the Chernoff criterion against other competing methods [12, 18, 33]. Since our purpose in this work is to provide theoretical insights into the Chernoff criterion, only a few experiments are performed here.

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. (1.1)), Chernoff criterion (Eq. (3.4)), and FKT [9].

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

(1) 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. (2) 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

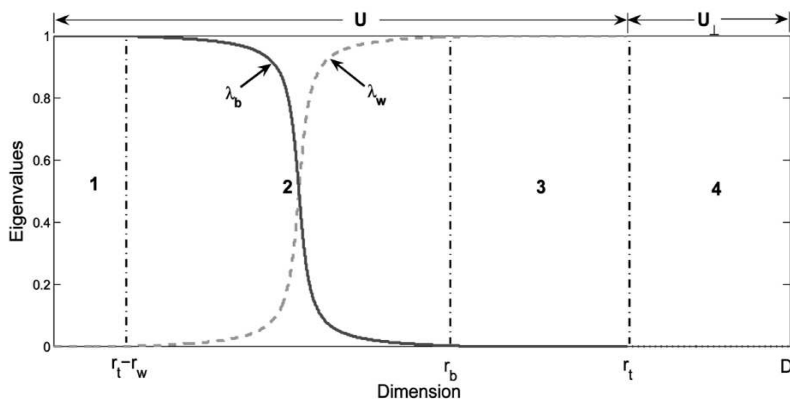


Figure 8: The entire data space is decomposed into four subspaces via FKT as described in [11]. There is no discriminant information in U_{\perp} , the null space of $S_b + S_w$. In U , $\lambda_b + \lambda_w = 1$.

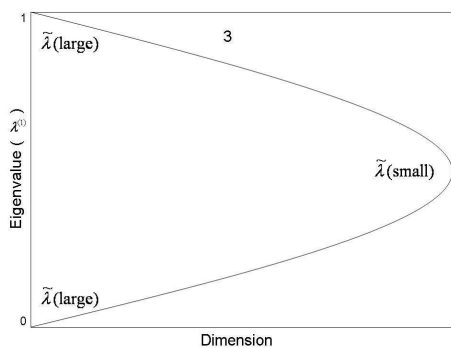


Figure 9: Augmented subspace 3.

instances after removing missing values. (3) 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. (4) 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 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. (5) 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. (6) 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. (7) 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. from each class. Thus, the data set consists of 1565 letter images. (8) 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. (9) 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.

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

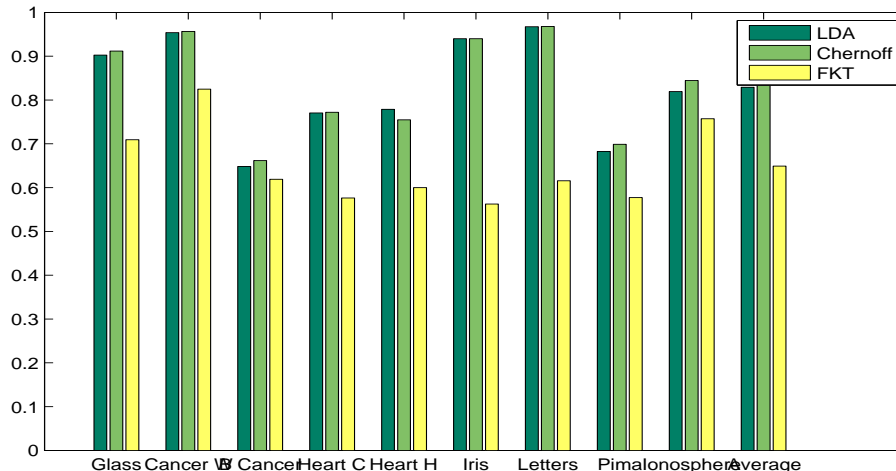


Figure 10: Classification error rates in subspaces computed by Fisher, Chernoff, and FKT using 1-nn classifier, on 11 UCI data sets.

sufficient.

Figure 10 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 [12, 18].

8 Summary

We note that the summary presented here also appeared in our earlier work [14].

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 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 has been 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 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] T. Cootes and C. Taylor, “Active shape models: Smart snakes,” in *Proc. British Machine Vision Conf.*, 1992, pp. 9–18.
- [3] S. Nayar, S. Baker, and H. Murase, “Parametric feature detection,” in *IEEE Conf. Computer Vision and Pattern Recognition*, 1994, pp. 471–477.
- [4] 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.
- [5] 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.
- [6] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [7] 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.
- [8] 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.

- [9] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] M. Loog, *Approximate Pairwise Accuracy Criteria for Multiclass Linear Dimension Reduction: Generalisations of the Fisher Criterion*. Delft Univ. Press, 1999.
- [14] J. Peng, S. Robila, W. Fan, and G. Seetharaman, "Analysis of chernoff criterion for linear dimensionality reduction," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 2010, pp. 3014–3021.
- [15] H. Decell and S. Mayekar, "Feature combinations and the divergence criterion," *Computers and Math. with Applications*, vol. 3, pp. 71–76, 1977.
- [16] P. Devijver and J. K. Pattern, *Recognition: A Statistical Approach*. London: Prentice-Hall, 1982.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] L. F. Chen, H. Y. Liao, M. T. Ko, J. C. Lin, and G. J. Yu, "A new lda-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, pp. 1713–1726, 2001.
- [21] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of lda," in *Proceedings of 16th International Conference on Pattern Recognition*, vol. 3, 2002, pp. 29–32.
- [22] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise lda for solving the small sample size problem in face verification," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 506–519, 2007.
- [23] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 157–165, January 2006.
- [24] J. Ye and Q. Li, "A two-stage linear discriminant analysis via qr-decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 929–942, 2005.
- [25] 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.
- [26] C. V. Loan, "Generalizing the singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 13, no. 1, pp. 76–83, 1976.
- [27] C. Paige and M. Saunders, "Towards a generalized singular value decomposition," *SIAM Journal on Numerical Analysis*, vol. 18, no. 3, pp. 398–405, 1981.
- [28] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of 9th International Conference on Machine Learning*, 1992, pp. 249–256.
- [29] Y. Sun and D. Wu, "A relief based feature extraction algorithm," in *Proceedings of SIAM International Conference on Data Mining*, 2008, pp. 188–195.
- [30] 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.
- [31] W. Bian and D. Tao, "Max-min distance analysis by using sequential sdp relaxation for dimension reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [32] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge, UK: Cambridge University Press, 2000.
- [33] D. Tao, X. Li, X. Wu, and S. Maybank, "Geometric mean for subspace selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 260–274, 2009.
- [34] C. Chen, "On information and distance measures, error bounds, and feature selection," *The Information Scientist*, vol. 10, pp. 159–173, 1979.
- [35] 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.
- [36] 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.