Approaches for Multi-Class Discriminant Analysis for Ranking Principal Components

Tiene Andre Filisbino
Laboratório Nacional de Computação Científica - LNCC
Petrópolis, RJ 25651-075
Email: tiene@lncc.br

Gilson Antonio Giraldi
Laboratório Nacional de Computação Científica - LNCC
Petrópolis, RJ 25651-075
Email: gilson@lncc.br

Carlos Eduardo Thomaz
Departamento de Engenharia Elétrica
Centro Universitário da FEI
São Bernardo do Campo, SP 09850-901
Email: cet@fei.edu.br

Abstract—The problem of ranking features computed by principal component analysis (PCA) in N-class problems have been addressed by the multi-class discriminant principal component analysis (MDPCA) and the Fisher discriminability criterion (FDC). These methods are motivated by the fact that PCA components do not necessarily represent important discriminant directions to separate sample groups. Given a database, the MDPCA builds a linear support vector machine (SVM) ensemble to get the separating hyperplanes that are combined through an AdaBoost technique to determine the discriminant contribution of each PCA feature. The FDC technique sorts PCA components according to the ratio of the between-class scatter over the within-class scatter. In this paper, we review these techniques and compare their performance in facial expression experiments. The classification results have shown the benefits of sorting principal components using FDC and the MDPCA though both methodologies are not so efficient when compared with PCA for reconstruction tasks.

I. INTRODUCTION

In pattern recognition applications, linear dimensionality reduction [15] and discriminant analysis are important steps for discarding redundancy and reduce the feature space for discriminating sample groups [5]. The principal component analysis (PCA) is a successful approach to the problem of creating a low dimensional data representation and interpretation of face images [15]. However, despite of the PCA capabilities, it was observed that, since PCA explains the covariance structure of all the data its most expressive components, that is, the first principal components with the largest eigenvalues, do not necessarily represent the most important discriminant directions to separate sample groups [11]. This observation motivates the development of new techniques to seek for discriminant PCA subspaces. In the context of two-class problems such solutions encompass Gaussian mixture [17], Zhu and Martinez technique as well as the discriminant principal components analysis (DPCA) method (see [11] and references therein).

In general, Fisher’s Linear Discriminant Analysis (LDA) [5] is used to compute the most important linear directions for separating sample groups [5], [2] rather than PCA. This method has the limitation of finding number of groups - 1 meaningful discriminant directions.

So, inspired in the LDA criterion, it is proposed in [18], [19] the Fisher discriminability criterion (FDC) that ranks PCA components according to the ratio of the between-class scatter over the within-class scatter. Also, the work [16] presents the multi-class discriminant principal components analysis (MDPCA), an extension of DPCA for N-class problems, that consists of the following steps: (a) Apply PCA technique for dimensionality reduction in order to eliminate redundancy, (b) Compute a linear SVM ensemble, based on the one-against-all SVM multi-class approach [1], (c) Use AdaBoost techniques [7] to combine the separating SVM hyperplanes in order to determine the global discriminant vector. It is important to highlight that both the MDPCA and FDC do not deal with the problem of computing general discriminant directions that are not principal components, like LDA does. Up to the best of our knowledge, MDPCA and FDC are the only techniques for ranking PCA components according to their discriminant power, for N-class problems with $N \geq 3$. This fact motivates our work to compare them.

So, in this paper we firstly review MDPCA and the FDC for ranking PCA components. Next, we evaluate these discriminant techniques for group separation tasks in facial expression experiments involving neutral, happiness, sad, anger, fear, frontal face images. Since the face image analysis problem involves large number of features and does not require a specific knowledge to interpret the differences between groups, it is an attractive application to investigate the effectiveness of the approaches for discriminant analysis in PCA spaces.

The classification experiments demonstrate the benefits of sorting principal components using FDC and the MDPCA if compared with the traditional PCA methodology of selecting as the principal components the ones with the largest eigenvalues. Moreover, the performed analysis indicates a superiority of FDC against the MDPCA for facial expression recognition tasks. However, the superiority of the FDC for classification is not observed for reconstruction tasks once it is outperformed by PCA and MDPCA methodologies. Besides, the PCA has lower reconstruction error when compared with MDPCA.

The paper is organized as follows: The theory behind PCA is presented in Section II. Next, in section III, the discriminant techniques of interest are revised. The computational experiments are described in Section IV. Finally, in Section V, we conclude the paper, summarizing its main contributions and describing further developments.
II. PRINCIPAL COMPONENTS ANALYSIS (PCA)

PCA is a feature extraction procedure that is related to the problem of efficiently representing the data, originally belonging to a \(m\)-dimensional space, through a \(m'\)-dimensional subspace where \(m' < m\). Hence, let an \(M \times m\) training set matrix \(\Theta\) be composed of \(M\) input samples (or face images), with \(m\) variables (or pixels), centralized respect to the global mean, computed by:

\[
\tilde{x} = \frac{1}{M} \sum_{i=1}^{M} x_i
\]  

(1)

Let this data matrix \(\tilde{\Theta}\) have covariance matrix \(\Omega\) with respectively \(P\) and \(\Lambda\) eigenvector and eigenvalue matrices, that is:

\[
P^T \Omega P = \Lambda.
\]  

(2)

It is a proven result that the set of \(m'\) \((m' \leq m)\) eigenvectors of \(\Omega\), which corresponds to the \(m'\) largest eigenvalues, minimizes the mean square reconstruction error over all choices of \(m'\) orthonormal basis vectors [5]. Such a set of eigenvectors that defines a new uncorrelated coordinate system for the training set matrix \(\tilde{\Theta}\) is known as the principal components. In the context of face recognition, those \(P_{\text{pca}} = \{p_1, p_2, \ldots, p_{m'}\}\) components are frequently called eigenfaces [12].

III. DISCRIMINANT ANALYSIS

It was observed that, since PCA explains the covariance structure of all the data its most expressive components [10], that is, the first principal components with the largest eigenvalues, do not necessarily represent important discriminant directions to separate sample groups.

Figure 1 is a simple example that helps to understand the limitation of PCA to select discriminant features for classification. Both Figures 1.(a) and 1.(b) represent the same data set. Figure 1.(a) just shows the PCA directions (\(\tilde{x}\) and \(\tilde{y}\)) and the distribution of the samples over the space. However, in Figure 1.(b) we distinguish two patterns: plus (+) and triangle (▼). We observe that the principal PCA direction \(\tilde{x}\) can not discriminate samples of the considered groups.

This observation motivates the application and development of techniques to compute discriminant subspaces, like FDC and MDPCA, which are described next.

A. Fisher Discriminability Criterion (FDC)

In the following sections we are supposing a data set \(X = \{(x_1, y_1), (x_2, y_2) \ldots (x_M, y_M)\}\), where \(x_i \in \mathbb{R}^m\), \(y_i \in \mathbb{Y}\) with \(Y = \{1, 2, 3, \ldots, N\}\). Following section II, the representation of data in the reduced space is given by:

\[
\tilde{x}_i = P^T \tilde{x}_i \in \mathbb{R}^{m'},
\]  

(3)

where \(\bar{x}_i = x_i - \bar{x}\) with, \(\bar{x}\) given by equation 1. The main purpose of Fisher’s Linear Discriminant Analysis (LDA) is to separate samples of distinct groups by maximizing their between-class separability while minimizing their within-class variability, by solving the problem [3], [8]:

\[
(P_{\text{Fisher}}) = \arg \max_P \frac{\sum_{c=1}^{C} N_c \cdot ||P^T \tilde{x}_c - P^T \tilde{x}||^2}{\sum_{i=1}^{M} ||P^T \tilde{x}_i - P^T \tilde{x}||^2}.
\]  

(4)

where \(C\) is the number of classes, \(N_c\) is the number of elements of class \(c\), \(\tilde{x}_c\) is the average of the samples belonging to class \(c\), \(\bar{x}\) is the average of all the samples and \(\tilde{x}_c\) is the average of the class corresponding to the \(i\)th sample.

Following, through expression (3), we can rewrite expression (4) as:

\[
(P_{\text{Fisher}}) = \arg \max_P \frac{\sum_{c=1}^{C} N_c \cdot \sum_{j} (\tilde{x}_{c,j} - \tilde{x})^2}{\sum_{i=1}^{M} \sum_{j} (\tilde{x}_{i,j} - \tilde{x})^2}.
\]  

(5)

By considering \(\|\cdot\|\) the usual 2-norm, we can interchange the summations to get:

\[
(P_{\text{Fisher}}) = \arg \max_P \sum_{j} \left( \frac{\sum_{c=1}^{C} N_c \cdot (\tilde{x}_{c,j} - \tilde{x})^2}{\sum_{i=1}^{M} (\tilde{x}_{i,j} - \tilde{x})^2} \right).
\]  

(6)

Expression (6) motivates the discriminability criterion presented in [19]. Specifically, from expression (6), we can postulate that the larger is the value of \(W_{j,\text{Fisher}}\) computed by:

\[
W_{j,\text{Fisher}} = \frac{\sum_{c=1}^{C} N_c \cdot (\tilde{x}_{c,j} - \tilde{x})^2}{\sum_{i=1}^{M} (\tilde{x}_{i,j} - \tilde{x})^2}.
\]  

(7)

then more discriminant is the \(p_j\) components for samples classification.

B. Multi-Class Discriminant Analysis

The Multi-Class DPCA technique [16], is based on the weakened SVM proposed in [4], the AdaBoost algorithm described in [14], page 25, and the AdaBoost.MH methodology presented in [9]. The Multi-Class DPCA is described by the Algorithm 2. The training instances in the input database \(X\) are supposed independent and identically distributed from an uniform distribution \(D_1\), at the initialization of the pipeline.
(line 1 of the Algorithm 2). Multi-Class DPCA also applies the PCA, revised on section II, in its first stage.

Next, the Multi-Class DPCA computes a SVM ensemble, based on the one-against-all SVM multi-class approach presented in [13]. So, let \( N \) be the total number of classes. Each iteration \( t \) of the Algorithm 2 constructs one weak SVM (line 7 of Algorithm 2), in the PCA subspace, using the Algorithm 1, presented in [4], and the training set built in line 6 of Algorithm 2.

\[ \text{Algorithm 1: WSVM Procedure: Build a Weakened version of SVM.} \]

**Input:** Labeled samples: \( X = \{ (x_i, y_i), i = 1, 2, \ldots, n' \} \) where \( y_i \in Y \) is the label of the sample \( x_i \);
Samples probability distribution \( D(x_i) \);
Percentage \( \mu \);
Select \( J \) so that, \( \sum_{j \in J} D(x_j) \leq (1 - \mu) \);
Select \( (x_i, y_i); i \in J \), and define \( D^* = D_{|J} \);
Compute the weighted data \( X^* = \{ (D^*_i \cdot x_i, y_i), i \in J \} \);
Compute the (weak) SVM hyperplane \( \phi_{\text{svm}} \) using \( X^* \);

**Output:** SVM hyperplane \( \phi_{\text{svm}} \).

Lines (8)-(12) of the Algorithm 2 are based on the AdaBoost philosophy [14]; in this case, to derive a strong classifier by using the linear combination of weak (SVM) learners \( h_1, h_2, \ldots, h_N \):

\[ H(x) = \sum_{t=1}^{N} \alpha_t h_t(x), \]

where \( \alpha_t \) does not depend on \( x \) [14].

The mathematical formulation of this idea is obtained by minimizing the corresponding exponential loss function and it gives rise to the expression in line 12 of the Algorithm 2 which updates the sample distribution \( D_t \) (see [14] for details).

The final step of the MDPCA procedure is justified by expression (8). In fact, if \( w_1, w_2, \ldots, w_N \) are the solution vectors that define the weak linear SVM classifiers then:

\[ H(x) = \sum_{t=1}^{N} \alpha_t w_t \cdot x, \]

where:

\[ < w_{\text{mdpca}}, x >= \sum_{t=1}^{N} \alpha_t w_t \cdot x = \sum_{i=1}^{m'} \left( \sum_{t=1}^{N} \alpha_t w_{t,i} \right) x_i, \]

So, following the same philosophy of the binary case [11], the Multi-Class DPCA determines the discriminant contribution of each feature by investigating the weights \( w_{\text{mdpca},i} = \sum_{t=1}^{N} \alpha_t w_{t,i} \). In fact, weights that are estimated to be 0 or approximately 0 have negligible contribution on the discriminant score \( H(x) \) given by equation (9), indicating that the corresponding features are not significant to separate the sample groups. In contrast, largest weights (in absolute values) indicate that the corresponding features contribute more to the discriminant score and consequently are important to characterize the differences between the groups. Therefore, the manner that AdaBoost combines the weak classifiers gives a straightforward way to compute the discriminant weights (line 13 of the Algorithm 2). In this way, instead of sorting features by selecting the corresponding principal components in decreasing order of eigenvalues, as PCA does, MDPCA selects as the most important features for classification the ones with the highest discriminant weights, that is, \( |w_{\text{mdpca},1}| \geq |w_{\text{mdpca},2}| \geq \cdots \geq |w_{\text{mdpca},m'}| \), as performed in lines 14-15 of the algorithm. The output of the MDPCA procedure is the discriminant principal components \( q_1, q_2, \ldots, q_{m'} \), where \( q_i \) is a PCA component ordered according to its discriminant weight \( |w_{\text{mdpca},i}| \).

**IV. Computational Experiments**

In this section, we perform experiments in facial expression classification and reconstruction using the frontal poses of the Radboud (RaFD) face image database which is an initiative of the Behavioural Science Institute of the Radboud University Nijmegen [6]. Like in [16], we take five different expressions: the neutral, happiness, sad, anger and fear frontal profile of
each person. In order to save memory allocation along the algorithms execution, we convert each pose to gray scale and resize it to 50 × 50 before computation.

Table I, lists the 10 principal components with the highest discriminant weights given by FDC (expression (7)) and MDPCA (Algorithm 2), in absolute values, for discriminating the expression samples in the following tasks:

- **Three-Class** experiment: happiness, neutral and sad samples;
- **Five-Class** experiment: happiness, neutral, sad, anger and fear classes.

In both experiments, three and five classes, we can observe that the FDC and MDPCA have selected some distant PCA components among the first 10 most discriminant principal ones. More specifically, we can observe that in the first 10, the first three most discriminant principal components selected by FDC and MDPCA are distant PCA components. We expect some consequences of this fact in the classification experiments, as we will see next.

---

### TABLE I: Top 10 discriminant principal components, ranked by FDC and the MDPCA procedure

<table>
<thead>
<tr>
<th>Expression Experiment: PCA components sorted by FDC and the MDPCA procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 classes FDC</td>
</tr>
<tr>
<td>3 classes MDPCA</td>
</tr>
<tr>
<td>5 classes FDC</td>
</tr>
<tr>
<td>5 classes MDPCA</td>
</tr>
</tbody>
</table>

---

To understand the changes described by the principal components, we reconstruct the most expressive features by varying each principal component \( p_i \), separately using the equation:

\[
I = \hat{x} + \beta \cdot p_i, \tag{10}
\]

where \( \hat{x} \) is given by expression (1), \( \beta \in \{\pm j \cdot \hat{\lambda}^{0.5}, j = 0, \pm 3\} \), and \( \hat{\lambda} \) is the average eigenvalue of the total covariance matrix \( \Omega \) in equation (2). We choose \( \hat{\lambda} \) instead of \( \lambda \) because some \( \lambda_i \) can be very small (or big) in this case, showing no changes (or color saturation) between the samples when we move along the corresponding principal components.

Figure 2 illustrates the transformations on the first PCA most expressive component contrasted with the first discriminant principal component selected by FDC and MDPCA to separate facial expressions. For Three-class and Five-class experiments, we can see that in the Figures 2.(a)-(c) and Figures 2.(j)-(l), respectively, the first PCA most expressive direction captures the changes in gender, which are the major variations of all the training samples.

However, when we compare these results with the ones reconstructed by FDC (Figures 2.(d)-(f) and 2.(m)-(o)) and MDPCA most discriminant principal components (Figures 2.(g)-(i) and 2.(m)-(o)), we can see that other principal components (the 24, 26, 27 according to Table I) carry more information about expression variations then the first PCA one. We expect some consequences of this fact in the recognition experiments. In fact, when PCA components are ranked by their largest discriminant weights rather than eigenvalues the result may be more effective with respect to extracting group-differences information as we can observe in Figure 3.

Figure 3 represents the value of the projection (score) of each sample in the principal component selected by PCA (Figure 3.(a)), FDC (Figure 3.(c)) and MDPCA (Figure 3.(e)). On the other hand, Figures 3.(b),(d) and Figure 3.(f) represent the samples projected on the first two principal components selected by the same methodologies, respectively.

Both Figures 3.(a)-(b) show the limitation of PCA to separate sample groups. In fact, from Figure 3.(a), if a sample has a high positive score it can belong to any one of the considered classes with (almost) the same probability. Analogous claim is true for negative scores. The samples distribution in Figure 3.(b) is somehow a consequence of this observation because the samples are mixed without some noticeable order. Figures 3.(c)-(d), obtained using FDC discriminant components, indicate a different behavior. In this case, Figure 3.(c) shows that samples with higher positive scores are more likely to belong to green class (happiness) than to the other ones. Figure 3.(d) agrees with this observation. When considering the scores for MDPCA (Figure 3.(e)) and the samples projection in Figure 3.(f) we can notice that MDPCA is more efficient to distinguish sample groups than PCA but its capability for
classification is inferior in the considered subspaces.

For Five-class experiments, Table I shows that the first and second principal components selected by the FDC and MDPCA are equal. So, we compute the scores using the 3th discriminant principal component selected by these methods. Figure 4 presents the obtained results. In this case, we also observe that the FDC (Figure 4.(a)) slightly outperforms MPCA (Figure 4.(b)) methodology because FDC can distinguish the green class (happiness) against magenta class (fear) with some efficiency and the scores for MDPCA (Figure 4.(b)) do not allow to visualize an analogous separation. However, from Figure 4.(a), we could not separate the blue, red and black classes from the green one. These results suggest that when increasing the number of classes the performance of FDC and MDPCA for separating samples groups decrease, as we will confirm next.

The following recognition tasks experiments are carried out using the full rank PCA subspace (section II) with all non-zero eigenvalues. We use the 10-fold cross validation method to evaluate the classification performance of PCA, FDC and the MDPCA techniques. In these experiments we have assumed equal prior probabilities and misclassification costs for all the classes. On the PCA subspace, the mean of each class \( i \) has been calculated from the corresponding training images and the Mahalanobis distance from each class mean \( \bar{x}_i \) has been used to assign a test observation \( x_r \) to one of the considered facial expressions. That is, we have assigned \( x_r \) to class \( i \) that minimizes:

\[
d_i(x_r) = \sum_{j=1}^{k} \left( \lambda_j (x_{rj} - \hat{x}_{ij})^2 \right),
\]

where \( \lambda_j \) is the corresponding eigenvalue, \( k \) is the number of principal components retained, \( x_{rj} \) and \( \hat{x}_{ij} \) are the projections of the sample \( x_r \) and of the mean \( \bar{x}_i \), respectively, in the \( j \)th component considered.

Figure 5 shows the average recognition rates of the 10-fold cross validation experiments for PCA, FDC and MDPCA, with Three-class represented by solid line and Five-class represented by dashed line. We can notice that, for the Three-class (solid line) recognition task the FDC and the MDPCA achieve higher recognition rates than the traditional PCA when considering \( k < 88 \). For \( 88 \leq k \leq 100 \) the recognition rates of PCA subspaces are higher or equal to the MDPCA methodology. For \( k \leq 100 \), FDC classification rates are higher than the other methodologies.

In same Figure 5 but for Five-class (dashed line), the recognition rates of FDC are higher or equal the rates of the other methods in \( 0 < k < 100 \). In range \( 70 < k \leq 100 \), PCA subspace outperforms the MDPCA. In \( k < 70 \), the PCA subspace has lower recognition rates than FDC and MDPCA methodologies. Moreover, in general, the recognition rates of FDC and MDPCA components are higher for Three-class than for Five-class tests. The results pictured on Figures 3 and 4 already pointed out the mentioned facts when using few components selected by the considered methods.

As already discussed in section III, since PCA explains features that most vary in the samples the principal subspaces do not necessarily represent important discriminant directions to separate sample groups. However, the reconstruction results are expected to give lower errors if we take components with higher variances. To make clear this observation, let us quantify the reconstruction quality through the root mean squared error (RMSE), computed as follows:

\[
RMSE^{ij}(k) = \sqrt{\frac{\sum_{i=1}^{N} ||P_i^j P^T x_i - \bar{x}_i||^2}{N}}.
\]
Average Recognition Rate

Fig. 5: Average recognition rate of PCA selected by the largest eigenvalues, FDC and largest MDPCA discriminant weights criteria, where three-class is solid line and Five-class is dashed line.

where \( i \in \{\text{PCA, FDC, MDPCA}\} \), \( i \in \{3, 5\} \) and \( I_k^i \) is a truncated identity matrix that keeps the subspace with dimension \( k \) that is selected by PCA, FDC or MDPCA.

Figure 6, we take the Five-class to show the RMSE for the reconstruction process for the subspaces given by the focused techniques. It is noticeable that PCA reconstruction performs better than the corresponding FDC and MDPCA discriminant components for all the simulated values of \( k \leq 194 \). Another point observed, for \( 14 < k < 170 \), the RMSE by MDPCA is better than FDC though in the same range FDC outperforms MDPCA in the classification task (solid lines in Figure 5).

More specifically, by merging the results observed in Figures 5 and 6, we conclude that the FDC is the best one for classification and PCA outperforms the other techniques in terms of reconstruction. However, in the range \( 14 < k < 70 \), MDPCA gives the recognition rates higher than PCA and it outperforms FDC in terms of reconstruction. Therefore, in applications that requires suitable (not the best) classification and good reconstruction the MDPCA can fulfill both requirements.

V. CONCLUSION

In this paper, we review MDPCA [16] and FDC [19] methodologies for ranking the PCA components computed using the Radboud facial expression database. Basically, we compute the discriminant weights for multi-class discriminant analysis using neural, happiness, sad, anger and fear frontal face images. The facial expressions experiments show that, in general, the principal components selected by FDC and MDPCA discriminant weights allow higher recognition rates using less linear features than standard PCA and that FDC outperforms MDPCA in this application for classification.