Uncorrelated Discriminant Projection Based on Maximum Margin Criterion and Its Kernelized Extension

Muhammad Amin¹ and Noor Atinah Ahmad¹
¹School of Mathematical Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia
muhammadaminu47@gmail.com, nooratinah@usm.my

Abstract. A new linear dimensionality reduction algorithm called uncorrelated discriminant projection (UDP) is proposed in this paper. The proposed UDP algorithm is based on the maximum margin criterion (MMC) which aim at maximizing class separation after dimension reduction. By imposing an uncorrelated constraint in the objective function, UDP extracts statistically uncorrelated features which are important in many pattern recognition problems. Moreover, we propose performing UDP in reproducing kernel Hilbert space (RKHS) which leads to a nonlinear variant of UDP called kernel uncorrelated discriminant projections (KUDP). In order to demonstrate the effectiveness and efficiency of the newly proposed algorithms, we conducted experiments on two benchmark face databases. The experimental results indicates that both UDP and KUDP are able to find face subspaces optimal for recognition.

1. Introduction

In many research areas such as computer vision and bioinformatics, data points are presented as points in high dimensional space. Researchers in such areas encounter difficulties working with such high dimensional data. The effectiveness and efficiency of many learning algorithms, such as clustering and classification algorithms, drops drastically when fed in high dimensional data. This undesirable property of high dimensional data is termed as “curse of dimensionality”. To deal with this problem, dimension reduction techniques are used to reduce the dimension of the data. The reduced data is then fed in to conventional learning algorithms.

One of the most widely used dimensionality reduction algorithm is linear discriminant analysis (LDA) [1]. LDA projects the high dimensional data to a lower-dimensional subspace in which data points from the same class are close to each other while data samples from different classes are mapped far apart. To further improve the discriminant ability of LDA, several variants of LDA such as regularized LDA [2], uncorrelated LDA [3], generalized uncorrelated LDA [4] and maximum margin criterion (MMC) [5] have been proposed. Uncorrelated LDA and generalized uncorrelated LDA are extensions of LDA in which a statistically uncorrelated constraint is added into the objective function of LDA in order to obtain uncorrelated discriminant features.

Uncorrelated discriminant feature extraction has received considerable interest in recent years [16–19]. The main idea of uncorrelated feature extraction is that features extracted using this approach are uncorrelated and contain minimum redundancy. Some of the most popular uncorrelated discriminant feature extraction techniques are uncorrelated LDA, uncorrelated discriminant locality preserving projections (UDLPP) [6] and uncorrelated discriminant analysis
(UDA) [7]. These methods have been shown to perform better than other methods which are not based on the same idea of statistically uncorrelation of feature vectors.

Motivated by the good performances of uncorrelated discriminant methods, we propose a new algorithm for linear dimensionality reduction called uncorrelated discriminant projections (UDP). Our approach essentially follows the same idea of MMC. By imposing a simple uncorrelated constraint, UDP can extracts uncorrelated features which are desirable in many pattern recognition problems. Specifically, features extracted using UDP are uncorrelated and contain minimum redundancy. Moreover, we extended the UDP algorithm to perform in RKHS which leads to a nonlinear variant of UDP called kernel uncorrelated discriminant projections (KUDP). We expect KUDP to perform better than UDP when the data is highly nonlinear.

The points below highlights some of the contributions of this paper:

- We provide a new formulation of the MMC optimization problem. Our proposed approach provide great possibilities that may lead to the development of new variants of MMC in the future.
- UDP is a linear dimensionality reduction method, which makes it suitable for practical applications. However, UDP can be carried out in RKHS which give rise to a nonlinear variant of UDP (KUDP).
- Both UDP and KUDP shares similar properties with the MMC algorithm. All methods aim at discovering the discriminant structure of the data. However, features extracted using UDP and KUDP are uncorrelated which is beneficial in many pattern recognition tasks.

The rest of the paper is structured as follows. We provide a brief review of the LDA and MMC algorithms in Section II. The newly proposed Uncorrelated Discriminant Projection algorithm is presented in Section III. In section IV, we extend the Uncorrelated Discriminant Projection algorithm to perform in RKHS. The experimental results on some benchmark face databases are given in section V. Some concluding remarks and future works are provided in section VI.

2. A brief review of LDA and MMC

In this section, we give a brief review of the LDA and MMC method. Suppose we have a set of \( n \) training samples \( X = [x_1, x_2, \ldots, x_n] \), belonging to \( C \) different classes, where \( x_i \in \mathbb{R}^m \) denotes the \( i \)-th sample. Let \( n_c \) be the number of samples in the \( c \)-th class, and \( \sum_{c=1}^{C} n_c = n \). In what follows, we give a brief review of the LDA and MMC algorithms.

2.1. LDA

LDA aims at finding a lower dimensional subspace in which data samples from the same remain close to each other while data samples from different class are mapped far apart. The LDA method solve the following objective function:

\[
 w^* = \arg \max_w \frac{w^T S_b w}{w^T S_w w}
\]  

\[
 S_b = \sum_{c=1}^{C} n_c (\mu^{(c)} - \mu)(\mu^{(c)} - \mu)^T,
\]  

\[
 S_w = \sum_{c=1}^{C} \sum_{j=1}^{n_c} (x^{(c)}_j - \mu^{(c)})(x^{(c)}_j - \mu^{(c)})^T,
\]
where \( \mu \) denotes the global centroid, \( \mu^{(c)} \) denotes the centroid of the \( c \)th class, \( n_c \) denotes the size of data samples in the \( c \)th class and \( x_j^{(c)} \) denotes the \( j \)th sample in the \( c \)th class.

When \( d \) projective functions \( W = [w_1, \ldots, w_d] \) are required, the objective function (1) can be reformulated as:

\[
W^* = \text{arg max}_W \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)}.
\]  

(4)

The objective function in (4) can be converted into the following generalized eigenproblem:

\[
S_b w = \lambda S_w w.
\]  

(5)

Thus the column vectors \( w_1, \ldots, w_d \) that form the transformation matrix \( W \) are computed as the eigenvectors corresponding to the first \( d \) largest eigenvalues of (5).

2.2. MMC

The matrix \( S_w \) in (5) is required to be nonsingular in order to obtain a stable solution of the eigenproblem. However, the matrix \( S_w \) tends to be singular when the number of features is larger than the number samples. To overcome this problem, a new method called maximum margin criterion (MMC) was introduced in [5]. The MMC method solved the following objective function:

\[
W^* = \text{arg max}_W \text{tr}(W^T (S_b - S_w) W).
\]  

(6)

The transformation matrix \( W = [w_1, \ldots, w_d] \) is required to comprised of unit vectors, i.e. \( w_k^T w_k = 1 \). Thus, the \( k \)th transformation vector \( w_k \) is obtained as the solution to the following objective function:

\[
w_k^* = \text{arg max}_{w_k} \frac{\text{tr}(S_b - S_w) w_k}{w_k^T w_k}.
\]  

(7)

The projection vectors \( w_k \) (for \( k = 1, \ldots, d \)) that maximizes (7) are calculate as the eigenvectors corresponding to the first \( d \) largest eigenvalues of the following eigenproblem:

\[
(S_b - S_w) w_k = \lambda_k w_k.
\]  

(8)

3. UDP

In order to introduce UDP, let us first reformulate MMC. Let \( \bar{x} = x_j - \mu \) denotes the centred data sample and \( \bar{X} = [\bar{x}_1, \ldots, \bar{x}_n] \) be the centred data matrix. The matrices \( S_b \) and \( S_w \) in MMC can be expressed as [8]:

\[
S_b = \bar{X} \bar{X}^T,
\]  

(9)

\[
S_w = \bar{X} L \bar{X}^T,
\]  

(10)

where \( L = I - S \) and \( S \) is defined as:

\[
S = \begin{bmatrix}
0 & \cdots & 0 \\
0 & S^{(2)} & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & S^{(c)}
\end{bmatrix},
\]  

(11)
where $\mathcal{S}_I$ is an $n_c \times n_c$ matrix having entries all equal to $1/n_c$. Using the above formulations, the objective function in (6) can be transformed into the following equivalent form:

$$W^* = \arg \max_{W} tr(W^T(\bar{X}M\bar{X}^T)W),$$

(12)

where $M = S.L = 2S1$. Thus, an optimal subspace for discriminant analysis can be obtained by solving the objective function in (12). The transformation matrix $W$ in MMC is sometimes required to be orthogonal. However, statistical uncorrelated feature have been shown to be superior to the orthogonality feature [9]. Statistically uncorrelated features contains minimum redundancy which is a favourable property in many pattern recognition problems.

Let $y_i$ and $y_j$ ($i = j$) be any two extracted features. These feature are said to be statistically uncorrelated if:

$$E[(y_i - E(y_i))(y_j - E(y_j))^T] = w_i^T S_i w_j = 0,$$

(13)

where $w_i$ and $w_j$ are two different columns of the transformation matrix $W$ and $S_i$ is the total scatter matrix of the training set. In order to simplify computations, $w_i$ is normalized to satisfy:

$$w_i^T S_i w_i = 1.$$

(14)

Equation (13) and (14) can be summarized as:

$$W^T S_i W = I.$$

(15)

Combining (12) and (15), we write the objective function of our proposed UDP method as:

$$W^* = \arg \max_{W^T S_i W = I} tr(W^T(\bar{X}M\bar{X}^T)W).$$

(16)

Since $S_i = \bar{X}\bar{X}^T$, we can write the objective function in (16) in the following equivalent form:

$$W^* = \arg \max_{W^T \bar{X}X^TW = I} tr(W^T(\bar{X}M\bar{X}^T)W).$$

(17)

The objective function in (17) can be transformed into the following generalized eigenproblem:

$$\bar{X}M\bar{X}^T w_k = \lambda_k \bar{X}\bar{X}^T w_k.$$

(18)

Let $w_1, \ldots, w_d$ be the solutions of (18) corresponding to the first $d$ largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. Then, the transformation matrix $W$ that optimizes (17) is given by $W = [w_1, \ldots, w_d]$. After $W$ is obtained, the uncorrelated linear features are extracted by $y = W^T \bar{x}$.

Note that the matrix $S_i = \bar{X}\bar{X}^T$ could sometimes be singular. We use the Tikhonov regularization technique to solve the singularity problem. The generalized eigenproblem in (18) then becomes:

$$\bar{X}M\bar{X}^T w_k = \lambda_k (\bar{X}\bar{X}^T + \gamma I) w_k,$$

(19)

where $\gamma > 0$ is the regularization parameter. The transformation matrix $W$ is obtained by solving (19).

It is also worth noting that, our proposed UDP algorithm is similar to the uncorrelated discriminant analysis (UDA) algorithm presented in [10]. However, there are some differences between the two approaches. The projection vectors $w_k$ that constitutes the transformation matrix $W = [w_1, \ldots, w_d]$ in UDA are chosen such that $w_i^T S_i w_k - w_i^T \bar{S}_b w_k = w_i^T \bar{X}M\bar{X}^T w_k > 0$.

Also, the discriminant information in the null space of $\tilde{S}_b$ ($\tilde{S}_b = W^T S_i W$) is thrown away in the UDA algorithm. Since the null space of $\tilde{S}_b$ might contain discriminant information [10], we keep the null space of $\tilde{S}_b$ for feature extraction because feature extraction is not only about reducing dimensionality but also retaining as much information in the original data as possible. Extending this argument, we also decided to relax the constraint $w^T \bar{X}M\bar{X}^T w_k > 0$. 


4. Kernel UDP

In this section, we extend UDP to perform in RKHS. Suppose the training data $X$ is mapped to a Hilbert space $H$ using a nonlinear mapping function $\phi : \mathbb{R}^m \rightarrow H$. Let $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$ denotes the training data matrix in $H$. The mapped training data vectors are centered such that $\bar{\phi}(x_i) = \phi(x_i) - \Phi_1$ where $\Phi_1$ is the global centroid defined as $\Phi_1 = \frac{1}{n} \sum_{i=1}^{n} \Phi_{i1}$, with $\Phi_{i1}$ being an $n$-element vector with entries all equal to one.

Now, the eigenproblem in $H$ can be written as follows:

$$\Phi M \Phi^T v_k = \lambda_k \Phi \Phi^T v_k. \tag{20}$$

Since the eigenvectors of (20) are linear combinations of $\bar{\phi}(x_1), \ldots, \bar{\phi}(x_n)$, there exist coefficients $a_i(i = 1, \ldots, n)$ such that

$$v_k = \sum_{i=1}^{n} a_i \bar{\phi}(x_i) = \bar{\Phi} a, \tag{21}$$

where $a = [a_1, \ldots, a_n]^T$. Simple algebra will allow us to rewrite (20) as:

$$\bar{K} M \bar{K} a = \lambda \bar{K} \bar{K} a, \tag{22}$$

where $\bar{K} = \bar{\Phi}^T \Phi = H \Phi^T \Phi H = HKH$. Moreover, the matrix $K = \Phi^T \Phi$ is called the kernel matrix and $\bar{K}$ is the centred kernel matrix. Let $a_1^*, \ldots, a^n$ be the $C - 1$ eigenvectors of the eigeproblem in (22). A new data point $x$ can be mapped into $C - 1$ dimensional subspace by:

$$x \rightarrow y = \Theta^T \bar{K}(\cdot, x), \tag{23}$$

where $\Theta = [a_1^*, \ldots, a^n]$ is the transformation matrix and $\bar{K}(\cdot, x) = \bar{\Phi}^T \bar{\phi}(x)$.

The matrix $\bar{K} \bar{K}$ is required to be nonsingular in order to obtained a stable solution of the eigenproblem in (22). In this paper, when $\bar{K} \bar{K}$ is singular, we use the Tikhonov regularization technique to solve the singularity problem. We then solve the following generalized eigenproblem instead:

$$\bar{K} M \bar{K} a = \lambda (\bar{K} \bar{K} + \gamma I) a. \tag{24}$$

where $\gamma > 0$ is the regularization parameter.

5. Experiments

In this section, we investigate the use of UDP and KUDP in face recognition. We compare the performances of these algorithms with that of the ordinary MMC algorithm on two widely used face databases, namely the CMU-PIE [11, 15] and AR database [12]. Different classifiers such as nearest neighbor classifier [1], random forest [13] and support vector machines [14] have been used in face recognition. In all our experiments, we used the nearest neighbor classifier with the Euclidean metric as our distance measure.

5.1. Face recognition on CMU-PIE database

The CMU PIE face database contains over 40,000 facial images of 68 individuals. The face images for each individual were captured across 13 different poses, under 43 different illumination conditions, and with 4 different expressions. A subset of the database which contains 2856 images of 68 individuals was used in our experiments. The facial areas were cropped and all images were

...
resized to $32 \times 2$ pixels. Thus, each image can be represented by a 1024-dimensional vector in image space. The features are scaled to [0, 1] in our experiment. Some sample images from the CMU-PIE database are shown in Fig. 1.

For each individual, $p (= 4, 6, 8, 10)$ images are randomly selected to form the training sets and the rest of the images are used as test sets. We average the result over 10 random splits for each $p$. Fig. 3 shows the average recognition rate versus the reduced dimensionality for each method. Also, the best average recognition rate, corresponding reduced dimensionality and standard deviation for each method are shown in Table 1.

As can be seen, UDP and KUDP significantly outperform the MMC method in terms of recognition accuracy. In particular, KUDP performed better than all methods for all $p$. However,
the UDP method performed comparatively to KUDP as p increases.

5.2. Face recognition on AR face database
The AR face database contains over 4000 color images of 126 individuals (70 men and 56 women). The images were taken in two different sessions (separated by two weeks). All images were taken using the same camera under different conditions of illumination, facial expressions and occlusions (sunglasses and scarf). Each image is of 768 x 576 pixels and each pixel is represented by 24 bits of RGB color values. Some samples images from the AR database are shown in Fig. 2.

We used a cropped subset of the AR database which contains 2600 images of 100 individuals in our experiment. There are 26 different images for each individual. We resized the images to 32 x 32 pixels and transformed them to grayscale. Each image is represented by a 1024-dimensional vector and the features (pixel values) are scaled to [0, 1]. We randomly choose p (= 4, 6, 8, 10) images per each individual to form the training sets and the rest of the images are considered as the test sets. We repeat this process 10 times and computed the average performance of each method. The best average recognition rate, corresponding dimensionality and standard deviations obtained by each method are reported in Table 2. We also plot the average recognition rates versus reduced dimensionality for the different methods in Fig. 4.

Figure 3: Recognition accuracy vs reduced dimensionality on CMU-PIE database (4, 6, 8 and 10 Train).
As can be seen, both UDP and KUDP outperform MMC, especially in small training sample case. This indicates that, both UDP and KUDP have more discriminating power than the MMC method and that the statistically uncorrelated property of extracted features is beneficial in improving recognition rate. Moreover, the optimal dimensionality obtained by KUDP is much lower than that obtained using UDP and MMC in almost all cases (4, 6, 8 and 10 Train).

6. Conclusion
A novel dimension reduction algorithm called uncorrelated discriminant projection (UDP) was proposed in this paper. UDP is based on the maximum margin criterion. Therefore, UDP has similar discriminant properties. Another property of the UDP algorithm is that the extracted features are uncorrelated which is a very important property in many pattern recognition tasks. We also propose a kernel variant of UDP in RKHS in order to handle highly nonlinear data. Performance improvement of our proposed algorithms over maximum margin criterion is demonstrated through face recognition experiments. Both UDP and KUDP were able to perform significantly better than MMC in all cases. In future work, we will examine the performances of the proposed methods on other pattern recognition tasks and consider using other classifiers different from the one used in this paper.
Acknowledgments
This work was supported by UNIVERSITI SAINS MALAYSIA through Research University Grant (RUI) (acc. No. 1001/PMATHS/8011040) and the USM Bridging Grant (acc. No. 304/PMATHS/6316049).

References
[1] Belhumeur PN, Hespanha JP, Kriegman DJ. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis & Machine Intelligence. 1997(7):711-20.
[2] Friedman JH. Regularized discriminant analysis. Journal of the American statistical association. 1989;84(405):165-75.
[3] Yuan D, Liang Y, Yi L, Xu Q, Kvalheim OM. Uncorrelated linear discriminant analysis (ULDA): A powerful tool for exploration of metabolomics data. Chemometrics and Intelligent Laboratory Systems. 2008;93(1):70-9.
[4] Ye J, Janardan R, Li Q, Park H. Feature reduction via generalized uncorrelated linear discriminant analysis. IEEE Transactions on Knowledge & Data Engineering. 2006(10):1312-22.
[5] Li H, Jiang T, Zhang K. Efficient and robust feature extraction by maximum margin criterion. Advances in neural information processing systems; 2004.
[6] Yu X, Wang X. Uncorrelated discriminant locality preserving projections. IEEE Signal Processing Letters. 2008;15:361-4.
[7] Ye J, Li T, Xiong T, Janardan R. Using uncorrelated discriminant analysis for tissue classification with gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2004;1(4):181-90.
[8] Cai D, He X, Han J. SRDA: An efficient algorithm for large-scale discriminant analysis. IEEE transactions on knowledge and data engineering. 2007;20(1):1-12.
[9] Yang J, Yang J-y, Zhang D. What’s wrong with Fisher criterion? Pattern recognition. 2002;35(11):2665-8.
[10] Yang W-H, Dai D-Q, Yan H. Feature extraction and uncorrelated discriminant analysis for high-dimensional data. IEEE transactions on knowledge and data engineering. 2008;20(5):601-14.
[11] Sim T, Baker S, Bsat M. The CMU pose, illumination, and expression (PIE) database. Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition; 2002:IEEE.
[12] Martínez AM, Kak AC. Pca versus lda. IEEE transactions on pattern analysis and machine intelligence. 2001;23(2):228-33.
[13] Zhang L, Suganthan PN. Random forests with ensemble of feature spaces. Pattern Recognition. 2014;47(10):3429-37.
[14] Guo G, Li SZ, Chan KL. Support vector machines for face recognition. Image and Vision computing. 2001;19(9-10):631-8.
[15] Aminu M, Ahmad NA. Locality preserving partial least squares discriminant analysis for face recognition. Journal of King Saud University-Computer and Information Sciences. 2019.
[16] Chen Z, Huang W, Lv Z. Towards a face recognition method based on uncorrelated discriminant sparse preserving projection. Multimedia Tools and Applications. 2017;76(17):17669-83.
[17] Sun S, Xie X, Yang M. Multiview uncorrelated discriminant analysis. IEEE transactions on cybernetics. 2015;46(12):3272-84.
[18] Ning X, Li W, Tang B, He H. BULDP: biomimetic uncorrelated locality discriminant projection for feature extraction in face recognition. IEEE Transactions on Image Processing. 2018;27(5):2575-86.
[19] Li X, Zhang H, Zhang R, Nie F. Discriminative and Uncorrelated Feature Selection with Constrained Spectral Analysis in Unsupervised Learning. IEEE Transactions on Image Processing. 2019.