Information Preserving Discriminant Projections

Jing Peng[1] and Alex J. Aved[2]
[1]*Department of Computer Science, Montclair State University, Montclair, NJ 07043, U.S.A.*
[2]*Information Directorate, AFRL, Rome, NY 13441, U.S.A.*

Keywords: Classification, Dimensionality Reduction, Feature Selection.

Abstract: In classification, a large number of features often make the design of a classifier difficult and degrade its performance. This is particularly pronounced when the number of examples is small relative to the number of features, which is due to the curse of dimensionality. There are many dimensionality reduction techniques in the literature. However, most these techniques are either informative (or minimum information loss), as in principal component analysis (PCA), or discriminant, as in linear discriminant analysis (LDA). Each type of technique has its strengths and weaknesses. Motivated by Gaussian Processes Latent Variable Models, we propose a simple linear projection technique that explores the characteristics of both PCA and LDA in latent representations. The proposed technique optimizes a regularized information preserving objective, where the regularizer is a LDA based criterion. And as such, it prefers a latent space that is both informative and discriminant, thereby providing better generalization performance. Experimental results based on a variety of data sets are provided to validate the proposed technique.

1 INTRODUCTION

In machine learning, a large number of features or attributes often make the development of classifiers difficult and degrade their performance. This is particularly pronounced when the number of examples is small relative to the number of features, which is due to the curse of dimensionality (Bellman, 1961). It simply states 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 d dimensional feature space requires estimating $O(2^d)$ unknown variables (Breiman et al., 1984).

Many machine learning problems are fundamentally related to the problem of learning latent representations or subspace learning, which potentially benefits many applications (Banerjee and Peng, 2005; Peng, 1995; Heisterkamp et al., 2000; **?**). The goal of subspace learning is to discover the geometric properties of the input space, such as its Euclidean embedding, intrinsic dimensionality, and connected components from a set of high dimensional examples. Subspace learning is also related to embedding. Subspace learning techniques can be categorized into linear and non-linear techniques. In this paper, we are mainly interested in linear techniques for simplicity and reduced computational complexity.

There are many dimensionality reduction techniques in the literature (Belhumeur et al., 1997; Fukunaga, 1990; Huo and et al, 2003; Howland and Park, 2004; Peng et al., 2013; Aved et al., 2017; Zhang et al., 2005). However, most these techniques are either informative (or minimum information loss), as in principal component analysis (PCA), or discriminant, as in linear discriminant analysis (LDA). Each type of technique has its strengths and weaknesses. For example, PCA is unsupervised, while LDA is supervised. Thus, it seems that in classification, LDA should be able to outperform PCA. However, it has been shown that PCA can outperform LDA in classification problems, given insufficient training data per subject (Martinez and Kak, 2001).

Motivated by Gaussian Processes (GP) latent variable models (Lawrence, 2005; Rasmussen and Williams, 2005; Urtasun and Darrell, 2007) and locality preserving projections (He and Niyogi, 2003; Cai et al., 2006), we investigate linear projection models that exploit the characteristics of PCA and LDA in latent representations. The latent representation computed by PCA is most informative in terms of minimum information loss. On the other hand, it does not take into account class label information. Thus,

Peng, J. and Aved, A.
Information Preserving Discriminant Projections.
DOI: 10.5220/0008967601620171
In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020) - Volume 2*, pages 162-171
ISBN: 978-989-758-395-7; ISSN: 2184-433X

it is less discriminant in general. By combining the characteristics of both PCA and LDA, it is expected that the resulting latent can be both informative and discriminant, thereby providing better generalization performance. Experimental results using a variety of data sets are provided to validate the proposed technique.

2 RELATED WORK

Many techniques have been proposed to take the advantage of the inherent low dimensional nature of the data (Darnell et al., 2017; Harandi et al., 2017; Sarveniazi, 2014; Xie et al., 2017). Two major linear subspace learning techniques are PCA and LDA. Both are capable of discovering the intrinsic geometry of the latent subspace. However, they only compute the global Euclidean structure.

Linear techniques based on graph Laplacians such as locality preserving projections (LPP) can model the local structure of the latent subspace (He and Niyogi, 2003). These techniques construct an adjacency matrix that captures the local geometry of the latent space from class label information. The projections are then computed by preserving such an adjacency structure. However, the basis functions obtained from LPP are not guaranteed to be orthogonal, which makes the data reconstruction more difficult.

Orthogonal LPP (OLPP) is a linear dimension reduction technique that has been proposed to address the problems associated with LPP (Cai et al., 2006). Similar to LPP, OLPP computes an adjacency matrix that preserves locality information. On the other hand, OLPP computes its basis functions that are guaranteed to be orthogonal, Orthogonal basis functions preserve the metric structure of the latent space. It is shown OLPP outperforms LPP (Cai et al., 2006).

GP latent variable models are probabilistic techniques for computing low dimensional subspaces from high dimensional data (Gao et al., 2011; Lawrence, 2005; Urtasun and Darrell, 2007; Jiang et al., 2012). These techniques have been applied to many problems such as image reconstruction and facial expression recognition (Abolhasanzadeh, 2015; Cai et al., 2016; Eleftheriadis et al., 2015; Song et al., 2015a).

GP latent variable models are generative and compute a latent subspace without taking into account class label information, as in (Lawrence, 2005). They are useful for visualization and regression analysis. And as such, the resulting latent space may not be optimal for classification. One way to address this problem is to introduce a prior distribution such as the uniform prior over the latent space to place constraints on the resulting latent space (Urtasun and Darrell, 2007). One of the main problems associated with GP latent variable models is that for a given test example, a separate estimation process must take a place to compute the corresponding latent position. This inference introduces additional uncertainties in the entire GP latent variable model computation and added computational complexity.

Integrating PCA and LDA for dimension reduction has been discussed in the literature (Yu et al., 2007; Zhao et al., 2011). These techniques formulate an objective as a linear combination of PCA and LDA criteria. On the other hand, we optimize the PCA objective with LDA as regularizer. This regularization view has a well established foundation in Gaussian Process latent variable models.

3 GAUSSIAN PROCESS LATENT VARIABLE MODELS

Gaussian Process (GP) latent variable models compute a low dimensional latent representation of high dimensional input data, using a GP mapping from the latent space to the input data space. Here we briefly describe GP latent variable models to motivate the introduction of the proposed technique.

Let

$$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^t$$

be a set of centered data, where $\mathbf{x}_i \in \Re^d$, and t denotes the transpose operator. Also, let

$$Z = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]^t$$

represent the corresponding latent variables, where $\mathbf{z}_i \in \Re^q$, and $q \ll d$. A typical relationship between the two sets of variables can be described by

$$\mathbf{x} = W\mathbf{z} + \varepsilon, \tag{1}$$

where W is a $d \times q$ matrix, and ε denotes the error term. Assuming that $p(\varepsilon) = N(0, \beta^{-1}\mathbf{I})$ (i.e., isotropic Gaussian), we have the following conditional probability distribution over the input space

$$p(\mathbf{x}|\mathbf{z}, W, \beta) = N(W\mathbf{z}, \beta^{-1}\mathbf{I}).$$

This implies that the likelihood of the data can be written as (matrix normal distribution)

$$p(X|Z, W, \beta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{z}_i, W, \beta).$$

Here it is assumed that $\mathbf{x}_i$ are independent and identically distributed (i.i.d.). Probabilistic PCA solution

for W can be computed by integrating out the latent variables (Tipping and Bishop, 1999).

A dual approach is to integrate out W and optimize the latent variables (Lawrence, 2005; Urtasun and Darrell, 2007). First, we specify a prior distribution $p(\mathbf{w}_i) = N(0, \alpha^{-1}\mathbf{I})$ for W, where $\mathbf{w}_i$ is the ith row of W. Then

$$p(W) = \prod_{i=1}^{d} p(\mathbf{w}_i) = \frac{1}{C_d}\exp(-\frac{1}{2}tr(\alpha W^t W)) = \frac{\alpha^{\frac{dq}{2}}}{(2\pi)^{\frac{dq}{2}}}\exp(-\frac{1}{2}tr(\alpha W^t W)). \quad (2)$$

where C_d is a normalization constant. Therefore, the marginalized likelihood of X can be computed by integrating out W

$$p(X|Z,\beta) = \int p(X|Z,W,\beta)p(W)dW \propto \frac{1}{|K|^{d/2}}\exp(-\frac{1}{2}tr(K^{-1}XX^t)), \quad (3)$$

where

$$K = (\alpha^{-1}ZZ^t + \beta^{-1}\mathbf{I}).$$

Thus, the distribution of the data given the latent variables is Gaussian. It can be shown that the solution Z, obtained by maximizing the GP likelihood of the latent variables (3), is equivalent to the PCA solution (Lawrence, 2005; Tipping and Bishop, 1999).

One can place additional conditions on the latent variables Z by introducing priors over Z. For example, if we place a uniform prior on Z, the log prior becomes

$$\ln p(Z) = -\frac{1}{2}\sum_{i=1}^{n}\mathbf{z}_i^t\mathbf{z}_i.$$

Such a prior prefers the latent variables close to the origin (Urtasun and Darrell, 2007). In classification context, one can incorporate class labels into the prior (Eleftheriadis et al., 2015; Song et al., 2015b). This can be accomplished based on discriminant analysis (Fukunaga, 1990). For example, LDA $J(Z) = tr(S_w^{-1}S_b)$, where S_w and S_b denote the between and within class matrices in the latent space, can be imposed. Here tr denotes the matrix trace. The prior thus becomes (Urtasun and Darrell, 2007)

$$p(Z) = C\exp(-J^{-1}).$$

One of drawbacks associated with GP latent models is that for a given test example, a separate estimation process must take a place to compute the corresponding latent position. This inference introduces additional uncertainties in the entire GP latent model computation and added computational complexity.

4 INFORMATION PRESERVING DISCRIMINANT PROJECTIONS

In this section, we develop a novel algorithm that combines some of the best features of Gaussian Process latent models and locality preserving projections.

As discussed above, the optimization of the likelihood (3) results in the PCA solution to the latent variables Z in the GP latent models. By introducing priors over latent variables $p(Z)$, one obtains the log posterior (terms that the posterior depends on)

$$L = -\frac{d}{2}\ln|K| - \frac{1}{2}tr(K^{-1}XX^t) + \ln p(Z). \quad (4)$$

As noted in (Rasmussen and Williams, 2005; Urtasun and Darrell, 2007), in a non-Bayesian setting, the log prior $\ln p(Z)$ can be viewed as a penalty term. And the maximum a posterior estimate of the latent variables can be interpreted as the penalized maximum likelihood estimate. If one introduces a discriminant prior, (4) represents a trade-off between informative (as in PCA) and discriminant (as in LDA) representations. A major drawback is that a separate estimation process must take place for each test example.

To overcome this problem, we introduce a simple linear projection technique that preserves the representation balance shown in (4), without a separate inference process for test examples. Recall that PCA finds the projection $\mathbf{p}$ by maximizing

$$J(\mathbf{p}) = tr(\mathbf{p}^t XX^t\mathbf{p}), \quad (5)$$

where XX^t denotes the covariance matrix, assuming the data are centered. Projection $\mathbf{p}$ has the property that the loss

$$\sum_{i}^{n}\|\mathbf{x}_i - \mathbf{p}\mathbf{p}^t\mathbf{x}_i\|^2$$

is minimum. Thus, PCA solutions are entirely informational in that the resulting latent representation preserves the maximum information.

To encourage the latent space to be discriminant, we appeal to the idea behind GP latent variable models, where the prior over the latent variables place constraints on variable positions in the latent space. As noted above, the log prior can be simply interpreted as a regularizer. Along this line, we can introduce a regularizer in (5)

$$J(\mathbf{p}) = tr(\mathbf{p}^t XX^t\mathbf{p}) + \lambda r(\mathbf{p}), \quad (6)$$

where $r(\cdot)$ represents a regularizer and λ is the regularization constant. $r(\cdot)$ plays the role of the log prior in GP latent models that places a constraint in the resulting latent space. In this work, we introduce the following regularizers: Laplacian and Linear Discriminant Analysis (LDA).

4.1 Information Preserving Projections with Laplacian Regularizer

A locality preserving projection (LPP) builds a graph of the input data that preserves local neighborhood information (He and Niyogi, 2003). LPP then computes a linear projection from the Laplacian of the graph.

Let W be a $n \times n$ weight matrix, where

$$W_{ij} = \begin{cases} \exp(-t\|\mathbf{x}_i - \mathbf{x}_j\|^2) & i \neq j \text{ and } l(\mathbf{x}_i) = l(\mathbf{x}_j) \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Here $\mathbf{x}_i$ represents the ith training example, $l(\cdot)$ denotes the label of its input, and t is a kernel parameter. Let $\mathbf{p}$ be a projection such that $z_i = \mathbf{p}^t\mathbf{x}_i$. LLP computes a linear projection by minimizing the following ojective

$$\sum_{i,j}(z_i - z_j)^2 W_{ij}.$$

A penalty is incurred when examples $\mathbf{x}_i$ and $\mathbf{x}_j$ that are in the same class are projected far apart. It turns out that the above objective can be rewritten as

$$\begin{aligned} \frac{1}{2}\sum_{i,j}(z_i - z_j)^2 W_{ij} &= \frac{1}{2}\sum_{i,j}(\mathbf{p}^t\mathbf{x}_i - \mathbf{p}^t\mathbf{x}_j)^2 W_{ij} \\ &= \mathbf{p}^t X^t L X \mathbf{p}, \end{aligned} \tag{8}$$

where $L = D - W$ is the graph Laplacian, and D is a diagonal matrix with diagonal entries $D_{ii} = \sum_j W_{ij}$. Since D_{ii} indicates the volume of z_i, LLP places the following constraint on the objective

$$\begin{aligned} &\min_{\mathbf{p}} \mathbf{p}^t X^t L X \mathbf{p} \\ \text{s.t.} \quad &\mathbf{p}^t X^t D X \mathbf{p} = 1 \end{aligned} \tag{9}$$

The optimal solution can be obtained by solving the generalized eigenvalue problem

$$X^t L X \mathbf{p} = \lambda X^t D X \mathbf{p}. \tag{10}$$

LPP has been shown to be effective in practice (He and Niyogi, 2003; Cai et al., 2006).

In LPP (He and Niyogi, 2003), the optimal projection $\mathbf{p}$ is computed by minimizing $\mathbf{p}^t X^t L X \mathbf{p}$. Therefore, the proposed Laplacian regularized PCA becomes

$$J(\mathbf{p}) = tr(\mathbf{p}^t X X^t \mathbf{p}) + \lambda tr(\mathbf{p}^t (X^t L X)^{-1} \mathbf{p}). \tag{11}$$

Thus, $\mathbf{p}$ can be computed by maximizing

$$J_{IP-Lap} = tr(X X^t + \lambda (X^t L X)^{-1}). \tag{12}$$

We call the resulting projection Information Preserving Laplacian Projection, or PLap. Note that (11) can be interpreted as a regularized PCA, where LPP is the regularizer. Or it can be viewed as a regularized LPP, where the regularizer is PCA.

4.2 Information Preserving Projections with LDA Regularizer

We can similarly introduce the LDA regularizer

$$J(\mathbf{p}) = tr(\mathbf{p}^t X X^t \mathbf{p}) + \lambda tr(\mathbf{p}^t S_w^{-1} S_b \mathbf{p}), \tag{13}$$

where

$$S_w = \sum_{c=1}^{C} \sum_{i=1, \mathbf{x}_i \in c}^{n_c} (\mathbf{x}_i - \mathbf{m}_c)(\mathbf{x}_i - \mathbf{m}_c)^t$$

and

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

are the between-class and within-class matrices. Here $\mathbf{m}$ represents the overall mean, $\mathbf{m}_c$ denotes the mean of class c, and t represents the transpose operator. Projection $\mathbf{p}$ can be computed by maximizing

$$J_{IP-LDA} = tr(X X^t + \lambda S_w^{-1} S_b). \tag{14}$$

We call the resulting projection Information Preserving LDA Projection, or PLda. Similar to PLap (11), (13) can be interpreted as a regularized PCA, where LDA is the regularizer. Or it can be viewed as a regularized LDA, where the regularizer is PCA.

5 EXPERIMENTS

We now examine the performance of the proposed techniques against competing techniques using several examples.

5.1 Methods

The following methods are evaluated in the experiments.

1. PLap–Information Preserving Projection with Laplacian regularizer (Eq. 12).
2. PLda–Information Preserving Projection with LDA regularizer (Eq. 14).
3. PCA–Projection that maximizes (Eq. 5)
$$J(\mathbf{p}) = tr(\mathbf{p}^t X X^t \mathbf{p}).$$
4. LDA–Projection that maximizes
$$J(\mathbf{p}) = tr((\mathbf{p}^t S_w \mathbf{p})^{-1} \mathbf{p}^t S_b \mathbf{p}),$$
where S_w and S_b are the within and between matrices, respectively.
5. OLPP–Orthogonal Laplacian Projection (OLPP) proposed in (Cai et al., 2006).

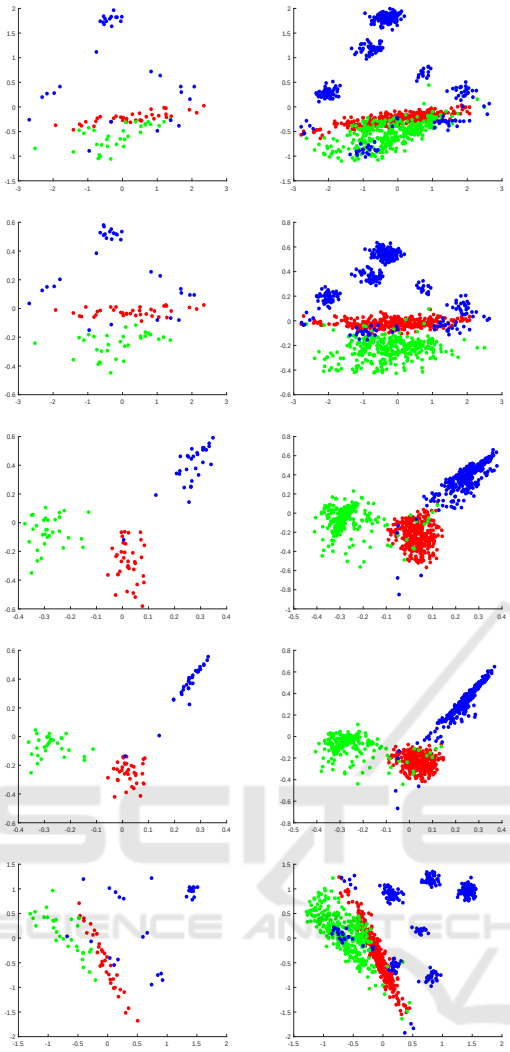

Figure 1: Oil flow data projected onto two dimensional spaces. The left column (top to bottom) shows the two dimensional representation of the training examples by PCA, LDA, PLap, PLda, and OLPP, respectively. The right column shows the two dimensional representation of the test examples by PCA, PLap, PLda, LDA, and OLPP, respectively.

Note that OLPP is along the line of LPP (9). It is known that the solution to (10) may not be orthogonal. To address this problem, OLPP first projects the data onto the PCA subspace, where it computes the solution to (10) that preserves orthogonality. It is shown in (Cai et al., 2006) OLPP outperforms LPP. Thus, we compare the proposed technique against OLPP in our experiments.

5.2 Oil Flow Data

We have carried out an experiment to visually examine our proposed techniques. The data set is the multiphase oil flow data (Bishop and James, 1993). This data set contains examples in 12 dimensions. The data set has three classes corresponding to the phase of flow in an oil pipeline: stratified, annular and homogeneous. In this experiment, for illustration purposes, we randomly sampled 100 examples as the training data, and additional 1000 examples as the test data.

Figure 1 shows the two dimensional projections by the five competing techniques. The left column (top to bottom) shows the two dimensional representation of the training examples by PCA, LDA, PLap, PLda, and OLPP, respectively. The right column shows the two dimensional representation of the test examples by the corresponding techniques. For PLap and PLda, regularization constant λ was to 100. Kernel parameter t in Laplacian (9) was set to 0.01 for both PLap and OLPP. The plots show that the proposed techniques provided better class separation than the competing methods in the latent space.

Figure 2 shows two dimensional projections of the test examples by PLap and PLda as a function of regularization constant λ: 20, 40, 60, and 80. The top panel shows the representation by PLap, and the bottom panel shows the representation by PLda. As λ increases, the resulting latent space becomes more discriminant. That is, as λ varies, the latent representations show the characteristics from PCA to LDA, as expected.

5.3 AR Face Data

This experiment involves the AR-face database (Martinez and Kak, 2001). The precise nature of the data set is described in (Martinez and Kak, 2001). For this experiment, 50 different subjects (25 males and 25 females) were randomly selected from this database. Images were normalized to the final 85×60 pixel arrays. Sample images are shown in Figure 3.

This experiment follows the exact setup of the Small Training Data set experiment described in (Martinez and Kak, 2001), where it is shown that PCA can outperform LDA, given insufficient training data per subject. Here we want to see how the proposed techniques fare against PCA in such situations.

In this setup, the first seven images from each subject are selected, resulting in a total of 350 images. To highlight the effects of a small training data set, two images from each subject are used as training and the remaining five are used as testing. Following (Martinez and Kak, 2001), we use all 21 different ways of

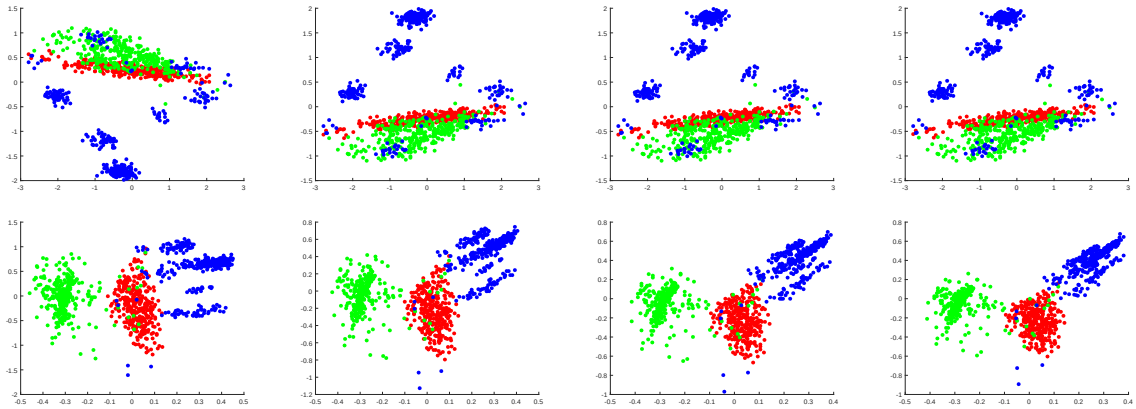

Figure 2: Two dimensional representation of the test examples by PLap and PLda as a function of regularization parameter λ: 20, 40, 60 and 80. The top panel shows the representation by PLap, and the bottom panel shows the representation by PLda. As λ increases, the resulting latent space becomes more discriminant.

Figure 3: AR sample images.

partitioning the data into training and testing for the results reported here. The original images of 85×60 pixels are first transformed via PCA into a space of 350 dimensions spanned by the 350 face data. All the methods see their input from this space.

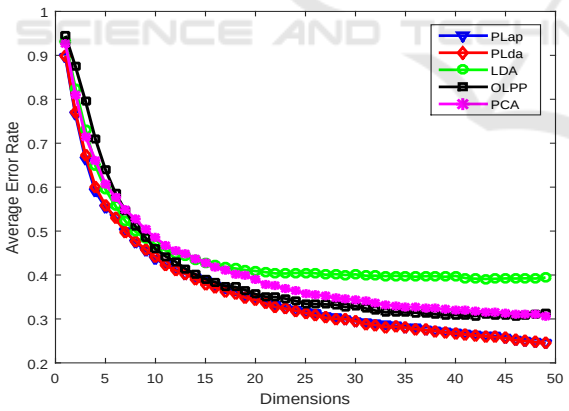

Figure 4: Average error rate registered by PLap, PLda, LDA, OLPP, and PCA with increasing dimensionality on the AR face data.

The average error rates in a subspace with 49 dimensions over the 21 runs are shown in the first row in Table 1. Figure 4 plots the average error rates registered by the competing methods over 21 runs on the AR face data as a function of increasing dimensions. On average both PLap and PLda clearly outperform PCA across the 49 subspaces, and consistently perform better than the competing methods.

5.4 Additional Data Sets

Additional data sets are used to illustrate the generalization performance by each competing technique.

1. *MNIST Data* (MNIST). The MNIST dataset consists of handwritten digits from the US National Institute of Standards and Technology (NIST) (yann.lecun.com/exdb/mnist/). Each digit is a 28 by 28 pixels of intensity values. Thus each digit is a feature vector of 784 intensities. In this experiment, we randomly selected 100 examples from each digit, for a totla of 1000 examples.
2. *Cat and Dog data* (CatDog). This image data set consists of two hundred images of cat and dog faces. Each image is a black-and-white 64×64 pixel image, and the images have been registered by aligning the eyes. Sample cat and dog images are shown in Figure 5.

Figure 5: Sample images of the cat and dog data.

3. *Multilingual Text* (MText). This data set is a multilingual text data set (Amini et al., 2009). It is from the Reuters RCV1 and RCV2 collections. The data set consists of six categories of documents: 1) *Economics*, 2) *Equity Markets*, 3) *Government Social*, 4) *Corporate/Industrial*, 5) *Performance*, and 6) *Government Finance*. Each document in English has a corresponding document

in French, German, Italian and Spanish, translated using PORTAGE (Ueffing et al., 2007). The documents in English are used in this experiment, where each document is represented by a bag of words model in 21531 dimensions. In this experiment, we randomly selected 100 examples from each class. Thus, the data set has 600 examples in 21531 dimensions.

4. *Iris Data* (Iris). The iris data set is a publicly available WVU multimodal data set (Crihalmeanu et al., 2007). The data set consists of iris images from subjects of different age, gender, and ethnicity, as described in (Crihalmeanu et al., 2007). The data set is difficult because many examples are low quality due to blur, occlusion, and noise. Sample iris images are shown in 6. The evaluation was done on a randomly selected pair of subjects, where one subject has 27 examples, and the other subject has 36 examples for a total of 63 examples.

   To compute features, iris images are segmented into 25×240 templates (Pundlik et al., 2008). Since Gabor features have been shown to produce better representation for iris data (Daugman, 2004), these templates are convolved with a log-Gabor filter at a single scale to obtain a 6000 features.

Figure 6: Sample Iris images.

5. *Fingerprint Data* (Finger). Similar to the iris data set, the finger data set is obtained from publicly available WVU multimodal data sets (Crihalmeanu et al., 2007). The data set consists of fingerprint images from a randomly chosen pair of subjects. Again, the data set is difficult because many examples are low quality due to blur, occlusion, and noise. Sample fingerprint images are shown in 7. This data set has 124 examples, where one subject has 61 instances, while the other has 63. To represent fingerprints, ridge and bifurcation features are computed using publically available code (sites.google.com/site/athisnarayanan/). Resulting feature vectors have 7241 components.

Figure 7: Sample fingerprint images.

6. *Feret Face Data* (FeretFace). The FERET face data set consists of 50 subjects, randomly chosen from the Feret face database (Phillips, 2004). Each subject has 8 instances. Therefore, we have a set of 400 facial images. The images used here involve variations in facial expressions and illumination. Each image has 150×130 pixels. Sample images are shown in Figure 8.

Figure 8: Normalized Feret sample images.

Table 1: Average error rates registered by the competing methods on the 7 data sets.

	PCA	PLap	PLda	LDA	OLPP
ARFace	0.307	0.245	0.245	0.394	0.314
MNIST	0.143	0.152	0.143	0.402	0.148
CatDog	0.492	0.216	0.210	0.457	0.286
MText	0.405	0.217	0.232	0.365	0.305
Iris	0.480	0.133	0.133	0.141	0.136
Finger	0.466	0.378	0.387	0.444	0.467
FeretFace	0.088	0.042	0.042	0.092	0.098
Ave	0.340	0.198	0.199	0.328	0.251

5.5 Experimental Results

In the additional six data set experiments, all training data have been normalized to have zero mean and unit variance along each dimension. The test data are similarly normalized using training mean and variance. In the resulting latent space, the one nearest neighbor rule is used to perform classification. The regularization constant λ and the kernel parameter t in Laplacian (9) were chosen through five fold cross validation. Table 1 shows the 10-fold crossed validated error rates of the five competing methods on the 6 data sets described above.

The table shows that on average both PLap and PLda outperformed all the competing techniques examined here. And PLap and PLda performed similarly on these data sets. The results show that classification performed in a space that is both informative and discriminant provides better generalization performance than in either the PCA or LDA subspace alone.

Figure 9 plots the 10-fold error rates computed by each method on the Feret face data as a function of increasing dimensions. As can be seen, both PLap and PLda consistently outperform the competing methods across the 49 subspaces, again demonstrating that classification performed in a space that is both informative and discriminant provides better generalization performance.

5.6 Robustness of Performance

PLap and PLda clearly achieved the best or near best performance over the 7 data sets, followed by OLPP. It seems natural to ask the question of robustness. That is, how well a particular method m performs on average in situations that are most favorable to other methods. We compute robustness by computing the ratio b_m of its error rate e_m and the smallest error rate over all methods being compared in a particular example:

$$b_m = e_m / \min_{1 \le k \le 5} e_k.$$

Thus, the best method m^* for that example has $b_{m^*} = 1$, and all other methods have larger values $b_m \geq 1$, for $m \neq m^*$. The larger the value of b_m, the worse the performance of the mth method is in relation to the best one for that example. The distribution of the b_m values for each method m over all the examples, therefore, seems to be a good indicator of robustness.

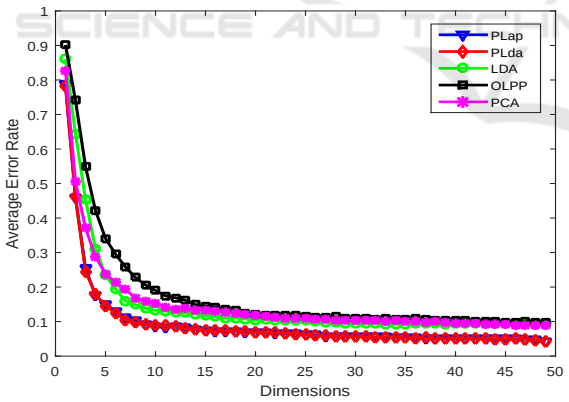

Figure 9: Average error rate registered by PLap, PLda, LDA, OLPP, and PCA with increasing dimensionality on the Feret face data.

Figure 10 plots the distribution of b_m for each method over the 7 data sets. The dark area represents the lower and upper quartiles of the distribution that are separated by the median. The outer vertical lines show the entire range of values for the distribution. It is clear that the most robust method over the data sets are PLap. In 5/7 of the data its error rate was the best (median = 1.0). In the worst case it was no worse than 62.9% higher than the best error rate. This is followed by PLda. In the worst case PLda was no worse than 69.1%. In contrast, PCA has the worst distribution, where the worst case was 360.9% higher than the best error rate.

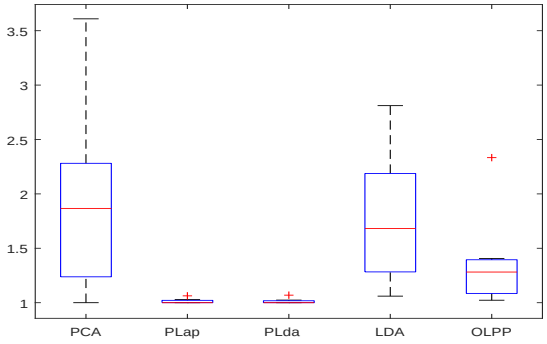

Figure 10: Error distributions of PCA, PLap, PLda, LDA, and OLPP over the 7 data sets.

6 SUMMARY

We have developed information preserving linear discriminant projections for computing latent representations. The proposed technique exploits the characteristics of PCA, LDA, and graph Laplacian to compute latent representations that are both informative and discriminant. As a result, the proposed technique provides better generalization performance. Experimental results are provided that validate the proposed technique. We note that the proposed technique is linear. We plan on extending this technique to the nonlinear case by incorporating kernel tricks in our future work.

REFERENCES

Abolhasanzadeh, B. (2015). Gaussian process latent variable model for dimensionality reduction in intrusion detection. In *2015 23rd Iranian Conference on Electrical Engineering*, pages 674–678.

Amini, M., Usunier, N., and Goutte, C. (2009). Learning from multiple partially observed views-an application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, pages 28–36.

Aved, A., Blasch, E., and Peng, J. (2017). Regularized difference criterion for computing discriminants for dimensionality reduction. *IEEE Transactions on Aerospace and Electronic Systems*, 53(5):2372–2384.

Banerjee, B. and Peng, J. (2005). Efficient learning of multi-step best response. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents*

and Multiagent Systems, AAMAS '05, pages 60–66, New York, NY, USA. ACM.

Belhumeur, V., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7):711–720.

Bellman, R. E. (1961). *Adaptive Control Precesses: A Guided Tour*. Princeton University Press.

Bishop, C. and James, G. (1993). Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 327(2):580–593.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Cai, D., He, X., Han, J., and Zhang, H. (2006). Orthogonal laplacianfaces for face recognition. *Trans. Img. Proc.*, 15(11):3608–3614.

Cai, L., Huang, L., and Liu, C. (2016). Age estimation based on improved discriminative gaussian process latent variable model. *Multimedia Tools Appl.*, 75(19):11977–11994.

Crihalmeanu, S., Ross, A., Schukers, S., and Hornak, L. (2007). A protocol for multibiometric data acquisition, storage and dissemination. In *Technical Report, WVU, Lane Department of Computer Science and Electrical Engineering*.

Darnell, G., Georgiev, S., Mukherjee, S., and Engelhardt, B. (2017). Adaptive randomized dimension reduction on massive data. *Journal of Machine Learning Research*, 18(140):1–30.

Daugman, J. (2004). How iris recognition works. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(21):21–30.

Eleftheriadis, S., Rudovic, O., and Pantic, M. (2015). Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 24(1):189–204.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Gao, X., Wang, X., Tao, D., and Li, X. (2011). Supervised gaussian process latent variable model for dimensionality reduction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(2):425–434.

Harandi, M., Salzmann, M., and Hartley, R. (2017). Joint dimensionality reduction and metric learning: A geometric take. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1404–1413, International Convention Centre, Sydney, Australia. PMLR.

He, X. and Niyogi, P. (2003). Locality preserving projections. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pages 153–160. MIT Press.

Heisterkamp, D., Peng, J., and Dai, H. (2000). Feature relevance learning with query shifting for content-based image retrieval. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 4, pages 250–253.

Howland, P. and Park, H. (2004). Generalizing discriminant analysis using the generalized singular valuke decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(8):995–1006.

Huo, X. and et al (2003). Optimal reduced-rank quadratic classiers using the fukunaga-koontz transform, with applications to automated target recognition. In *Proc. of SPIE Conference*.

Jiang, X., Gao, J., Wang, T., and Zheng, L. (2012). Supervised latent linear gaussian process latent variable model for dimensionality reduction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(6):1620–1632.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816.

Martinez, A. M. and Kak, A. (2001). Pca versus lda. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):228–233.

Peng, J. (1995). Efficient memory-based dynamic programming. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 438–446.

Peng, J., Seetharaman, G., Fan, W., and Varde, A. (2013). Exploiting fisher and fukunaga-koontz transforms in chernoff dimensionality reduction. *ACM Transactions on Knowledge Discovery from Data*, 7(2):8:1–8:25.

Phillips, P. (2004). The facial recognition technology (feret) database. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22.

Pundlik, S., Woodard, D., and Birchfield, S. (2008). Non-ideal iris segmentation using graph cuts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6.

Rasmussen, C. and Williams, C. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Sarveniazi, A. (2014). An actual survey of dimensionality reduction. *American Journal of Computational Mathematics*, 4(2):55–72.

Song, G., Wang, S., Huang, Q., and Tian, Q. (2015a). Similarity gaussian process latent variable model for multi-modal data analysis. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4050–4058.

Song, G., Wang, S., Huang, Q., and Tian, Q. (2015b). Similarity gaussian process latent variable model for multi-modal data analysis. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4050–4058.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 61(3):611–622.

Ueffing, N., Simard, M., Larkin, S., and Johnson, J. (2007). NRC's PORTAGE system for WMT 2007. In *In ACL-2007 Second Workshop on SMT*, pages 185–188.

Urtasun, R. and Darrell, T. (2007). Discriminative gaussian process latent variable model for classification. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 927–934, New York, NY, USA. ACM.

Xie, P., Deng, Y., Zhou, Y., Kumar, A., Yu, Y., Zou, J., and Xing, E. (2017). Learning latent space models with angular constraints. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3799–3810, International Convention Centre, Sydney, Australia. PMLR.

Yu, J., Tian, Q., Rui, T., and Huang, T. S. (2007). Integrating discriminant and descriptive information for dimension reduction and classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):372–377.

Zhang, P., Peng, J., and Domeniconi, C. (2005). Kernel pooled local subspaces for classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):489–502.

Zhao, N., Mio, W., and Liu, X. (2011). A hybrid pca-lda model for dimension reduction. In *The 2011 International Joint Conference on Neural Networks*, pages 2184–2190.