Pattern Recognition 42 (2009) 764 – 779

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Perturbation LDA: Learning the difference between the class empirical mean and its expectation

Wei-Shi Zheng[a,c], J.H. Lai[b,c,*], Pong C. Yuen[d], Stan Z. Li[e]

[a] *School of Mathematics and Computational Science, Sun Yat-sen University, Guangzhou, PR China*
[b] *Department of Electronics and Communication Engineering, School of Information Science and Technology, Sun Yat-sen University, Guangzhou, PR China*
[c] *Guangdong Province Key Laboratory of Information Security, PR China*
[d] *Department of Computer Science, Hong Kong Baptist University, Hong Kong*
[e] *Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, PR China*

ARTICLE INFO

Article history:
Received 24 September 2006
Received in revised form 9 July 2008
Accepted 22 September 2008

Keywords:
Fisher criterion
Perturbation analysis
Face recognition

ABSTRACT

Fisher's linear discriminant analysis (LDA) is popular for dimension reduction and extraction of discriminant features in many pattern recognition applications, especially biometric learning. In deriving the Fisher's LDA formulation, there is an assumption that the class empirical mean is equal to its expectation. However, this assumption may not be valid in practice. In this paper, from the "perturbation" perspective, we develop a new algorithm, called perturbation LDA (P-LDA), in which perturbation random vectors are introduced to learn the effect of the difference between the class empirical mean and its expectation in Fisher criterion. This perturbation learning in Fisher criterion would yield new forms of within-class and between-class covariance matrices integrated with some perturbation factors. Moreover, a method is proposed for estimation of the covariance matrices of perturbation random vectors for practical implementation. The proposed P-LDA is evaluated on both synthetic data sets and real face image data sets. Experimental results show that P-LDA outperforms the popular Fisher's LDA-based algorithms in the undersampled case.

1. Introduction

Data in some applications such as biometric learning are of high dimension, while available samples for each class are always limited. In view of this, dimension reduction is always desirable, and at the same time it is also expected that data of different classes can be more easily separated in the lower-dimensional subspace. Among the developed techniques for this purpose, Fisher's linear discriminant analysis (LDA)[1] [1–4] has been widely and popularly used as a powerful tool for extraction of discriminant features. The basic principle of Fisher's LDA is to find a projection matrix such that the ratio between the between-class variance and within-class variance is maximized in a lower-dimensional feature subspace.

Due to the curse of high dimensionality and the limit of training samples, within-class scatter matrix $\mathbf{S}_W$ is always singular, so that classical Fisher's LDA will fail. This kind of singularity problem is always called the small sample size problem [5,6] in Fisher's LDA. So far, some well-known variants of Fisher's LDA have been developed to overcome this problem. Among them, Fisherface (PCA+LDA) [5], nullspace LDA (N-LDA) [6–8] and regularized LDA (R-LDA) [9–13] are three representative algorithms. In "PCA+LDA", Fisher's LDA is performed in a principal component subspace, in which within-class covariance matrix will be of full rank. In N-LDA, the nullspace of within-class covariance matrix $\mathbf{S}_W$ is first extracted, and then data are projected onto that subspace and finally a discriminant transform is found there for maximization of the variance among between-class data. In R-LDA, a regularized term, such as $\lambda \cdot \mathbf{I}$ where $\lambda > 0$, is added to $\mathbf{S}_W$. Some other approaches, such as Direct LDA [14], LDA/QR [15] and some constrained LDA [16,17], are also developed. Recently, some efforts are made for development of two-dimensional LDA techniques (2D-LDA) [18–20], which perform directly on matrix-form data. A recent study [21] conducts comprehensive theoretical and experimental comparisons between the traditional Fisher's LDA techniques and some representative 2D-LDA algorithms in the undersampled case. It is experimentally shown

* Corresponding author at: Department of Electronics and Communication Engineering, School of Information Science and Technology, Sun Yat-sen University, Guangzhou, Guangdong 510275, PR China. Tel.: +86 020 84035440.
E-mail addresses: wszheng@ieee.org (W.-S. Zheng), stsljh@mail.sysu.edu.cn (J.H. Lai), pcyuen@comp.hkbu.edu.hk (Pong C. Yuen), szli@nlpr.ia.ac.cn (Stan Z. Li).

[1] LDA in this paper refers to Fisher's LDA. It is not a classifier but a feature extractor learning low-rank discriminant subspace, in which any classifier can be used to perform classification.

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.09.012

that some two-dimensional LDA may perform better than Fisherface and some other traditional Fisher's LDA approaches in some cases, but R-LDA always performs better. However, estimation of the regularized parameter in R-LDA is hard. Though cross-validation (CV) is popularly used, it is time consuming. Moreover, it is still hard to fully interpret the impact of this regularized term.

From the geometrical view, Fisher's LDA makes different class means scatter and data of the same class close to their corresponding class means. However, since the number of samples for each class is always limited in some applications such as biometric learning, the estimates of class means are not accurate, and this would degrade the power of Fisher criterion. To specify this problem, we first re-visit the derivation of Fisher's LDA. Consider the classification problem of L classes $C_1,\ldots,C_L$. Suppose the data space $\mathbf{X}$ ($\subset \Re^n$) is a compact vector space and $\{(\mathbf{x}_1^1,y_1^1),\ldots,(\mathbf{x}_{N_1}^1,y_{N_1}^1),\ldots,(\mathbf{x}_1^L,y_1^L),\ldots,(\mathbf{x}_{N_L}^L,y_{N_L}^L)\}$ is a set of finite samples. All data $\mathbf{x}_1^1,\ldots,\mathbf{x}_{N_1}^1,\ldots,\mathbf{x}_1^L,\ldots,\mathbf{x}_{N_L}^L$ are *iid*, and $\mathbf{x}_i^k$ ($\in \mathbf{X}$) denotes the ith sample of class C_k with class label y_i^k (i.e., $y_i^k = C_k$) and N_k is the number of samples of class C_k. The *empirical mean* of each class is then given by $\hat{\mathbf{u}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{x}_i^k$ and the total sample mean is given by $\hat{\mathbf{u}} = \sum_{k=1}^{L}\frac{N_k}{N}\hat{\mathbf{u}}_k$, where $N = \sum_{k=1}^{L}N_k$ is the number of total training samples. The goal of LDA under Fisher criterion is to find an optimal projection matrix by optimizing the following Eq. (1):

$$\hat{\mathbf{W}}_{opt} = arg\max_{\mathbf{W}} trace(\mathbf{W}^{\mathrm{T}}\hat{\mathbf{S}}_b\mathbf{W})/trace(\mathbf{W}^{\mathrm{T}}\hat{\mathbf{S}}_w\mathbf{W}), \tag{1}$$

where $\hat{\mathbf{S}}_b$ and $\hat{\mathbf{S}}_w$ are between-class covariance (scatter) matrix and within-class covariance (scatter) matrix, respectively, defined as follows:

$$\hat{\mathbf{S}}_b = \sum_{k=1}^{L}\frac{N_k}{N}(\hat{\mathbf{u}}_k - \hat{\mathbf{u}})(\hat{\mathbf{u}}_k - \hat{\mathbf{u}})^{\mathrm{T}}, \tag{2}$$

$$\hat{\mathbf{S}}_w = \sum_{k=1}^{L}\frac{N_k}{N}\hat{\mathbf{S}}_k, \quad \hat{\mathbf{S}}_k = \sum_{i=1}^{N_k}\frac{1}{N_k}(\mathbf{x}_i^k - \hat{\mathbf{u}}_k)(\mathbf{x}_i^k - \hat{\mathbf{u}}_k)^{\mathrm{T}}. \tag{3}$$

It has been proved in [22] that Eq. (2) could be written equivalently as follows:

$$\hat{\mathbf{S}}_b = \frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{L}\frac{N_k}{N}\times\frac{N_j}{N}(\hat{\mathbf{u}}_k - \hat{\mathbf{u}}_j)(\hat{\mathbf{u}}_k - \hat{\mathbf{u}}_j)^{\mathrm{T}}. \tag{4}$$

For formulation of Fisher's LDA, two basic assumptions are always used. First, the class distribution is assumed to be Gaussian. Second, the class empirical mean is in practice used to approximate its expectation. Although Fisher's LDA has been getting its attraction for more than thirty years, as far as we know, there is little research work addressing the second assumption and investigating the effect of the difference between the class empirical mean and its expectation value in Fisher criterion. As we know, $\hat{\mathbf{u}}_k$ is the estimate of $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ based on the maximum likelihood criterion, where $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ is the *expectation* of class C_k. The substitution of expectation $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ with its empirical mean $\hat{\mathbf{u}}_k$ is based on the assumption that the sample size for estimation is large enough to reflect the data distribution of each class. Unfortunately, this assumption is not always true in some applications, especially the biometric learning. Hence the impact of the difference between those two terms should not be ignored.

In view of this, this paper will study the effect of the difference between the class empirical mean and its expectation in Fisher criterion. We note that such difference is almost impossible to be specified, since $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ is usually hard (if not impossible) to be determined. Hence, from the "perturbation" perspective, we introduce the perturbation random vectors to stochastically describe such difference. Based on the proposed perturbation model, we then analyze how perturbation random vectors take effect in Fisher criterion. Finally, perturbation learning will yield new forms of within-class and between-class covariance matrices by integrating some perturbation factors, and therefore a new Fisher's LDA formulation based on these two new estimated covariance matrices is called *perturbation LDA* (P-LDA). In addition, a semi-perturbation LDA, which gives a novel view to R-LDA, will be finally discussed.

Although there are some related work on covariance matrix estimation for designing classifier such as RDA [23] and its similar work [24], and EDDA [25], however, the objective of P-LDA is different from theirs. RDA and EDDA are not based on Fisher criterion and they are classifiers, while P-LDA is a feature extractor and does not predict class label of any data as output. P-LDA would exact a subspace for dimension reduction but RDA and EDDA do not. Moreover, the perturbation model used in P-LDA has not been considered in RDA and EDDA. Hence the methodology of P-LDA is different from the ones of RDA and EDDA. This paper focuses on Fisher criterion, while classifier analysis is beyond our scope. To the best of our knowledge, there is no similar work addressing Fisher criterion using the proposed perturbation model.

The remainder of this paper is outlined as follows. The proposed P-LDA will be introduced in Section 2. The implementation details will be presented in Section 3. Then P-LDA is evaluated using three synthetic data sets and three large human face data sets in Section 4. Discussions and conclusion of this paper are then given in Sections 5 and 6, respectively.

2. P-LDA: a new formulation

The proposed method is developed based on the idea of perturbation analysis. A theoretical analysis is given and a new formulation is proposed by learning the difference between the class empirical mean and its expectation as well as its impact to the estimation of covariance matrices is Fisher criterion. In Section 2.1, we first consider the case when data of each class follow single Gaussian distribution. The theory is then extended to the mixture of Gaussian distribution case and reported in Section 2.2. The implementation details of the proposed new formulation will be given In Section 3.

2.1. P-LDA under single Gaussian distribution

Assume data of each class are normally distributed. Given a specific input $(\mathbf{x},y)$, where sample $\mathbf{x} \in \mathbf{X}$ and class label $y \in \{C_1,\ldots,C_L\}$, we first try to study the difference between a sample $\mathbf{x}$ and $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$ the expectation of class y in Fisher criterion. However, $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$ is usually hard (if not impossible) to be determined, so it may be impossible to specific such difference. Therefore, our strategy is to stochastically characterize (simulate) the difference between $\mathbf{x}$ and $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$ by a random vector and then model a random mean for class y to stochastically describe $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$. Define $\xi_{\mathbf{x}}$ ($\in \Re^n$) as a *perturbation random vector* for stochastic description (simulation) of the difference between $\mathbf{x}$ and $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$. When data of each class follow normal distribution, we can model $\xi_{\mathbf{x}}$ as a random vector from the normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Omega}_y$, i.e.,

$$\xi_{\mathbf{x}} \sim \mathbf{N}(\mathbf{0},\mathbf{\Omega}_y), \quad \mathbf{\Omega}_y \in \Re^{n\times n}. \tag{5}$$

We call $\mathbf{\Omega}_y$ the *perturbation covariance matrix* of $\xi_{\mathbf{x}}$. The above model assumes that the covariance matrices $\mathbf{\Omega}_y$ of $\xi_{\mathbf{x}}$ are the same for any sample $\mathbf{x}$ with the same class label y. Note that it would be natural that an ideal value of $\mathbf{\Omega}_y$ can be the expected covariance matrix of class y, i.e., $\mathbf{E}_{\mathbf{x}'|y}[(\mathbf{x}' - \mathbf{E}_{\mathbf{x}''|y}[\mathbf{x}''])(\mathbf{x}' - \mathbf{E}_{\mathbf{x}''|y}[\mathbf{x}''])^{\mathrm{T}}]$. However, this value is usually hard to be determined, since $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$ and the true density function are not available. Actually this kind of estimation needs not be our goal. Note that the perturbation random vector $\xi_{\mathbf{x}}$ is only used for stochastic simulation of the difference between the specific sample $\mathbf{x}$ and its expectation $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$. Therefore, in our study, $\mathbf{\Omega}_y$ only needs to be properly estimated for performing such simulation based on the perturbation model specified by the following Eqs. (6) and (7), finally resulting in some proper correctings (perturbations) on the empirical between-class and within-class covariance matrices as shown later. For this goal, a random vector is first formulated for any sample $\mathbf{x}$ to stochastically approximate $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$ below:

$$\tilde{\mathbf{x}} = \mathbf{x} + \xi_{\mathbf{x}}. \tag{6}$$

The stochastic approximation of $\tilde{\mathbf{x}}$ to $\mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']$ means there exists a specific estimate[2] $\hat{\xi}_{\mathbf{x}}$ of the random vector $\xi_{\mathbf{x}}$ with respect to the corresponding distribution such that

$$\mathbf{x} + \hat{\xi}_{\mathbf{x}} = \mathbf{E}_{\mathbf{x}'|y}[\mathbf{x}']. \tag{7}$$

Formally we call equality Eqs. (6) and (7) the *perturbation model*. It is not hard to see such perturbation model is always satisfied. The main problem is how to model $\mathbf{\Omega}_y$ properly. For this purpose, a technique will be suggested in the next section.

Now, for any training sample $\mathbf{x}_i^k$, we could formulate its corresponding perturbation random vector $\xi_i^k \sim \mathbf{N}(\mathbf{0}, \mathbf{\Omega}_{C_k})$ and the random vector $\tilde{\mathbf{x}}_i^k = \mathbf{x}_i^k + \xi_i^k$ to stochastically approximate its expectation $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$. By considering the perturbation impact, $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ could be stochastically approximated on average by:

$$\tilde{\mathbf{u}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \tilde{\mathbf{x}}_i^k = \hat{\mathbf{u}}_k + \frac{1}{N_k}\sum_{i=1}^{N_k} \xi_i^k. \tag{8}$$

Note that $\tilde{\mathbf{u}}_k$ can only stochastically but not exactly describe $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$, so it is called the *random mean* of class C_k in our study.

After introducing the random mean of each class, a new form of Fisher's LDA is developed below by integrating the factors of the perturbation between the class empirical mean and its expectation into the supervised learning process, so that new forms of the between-class and within-class covariance matrices are obtained. Since $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{u}}$ are both random vectors, we take the expectation with respect to the probability measure on their probability spaces, respectively. To have a clear presentation, we denote some sets of random vectors as $\xi^k = \{\xi_1^k, \ldots, \xi_{N_k}^k\}$, $k = 1, \ldots, L$, and $\xi = \{\xi_1^1, \ldots, \xi_{N_1}^1, \ldots, \xi_1^L, \ldots, \xi_{N_L}^L\}$. Since $\mathbf{x}_1^1, \ldots, \mathbf{x}_{N_1}^1, \ldots, \mathbf{x}_1^L, \ldots, \mathbf{x}_{N_L}^L$ are *iid*, it is reasonable to assume that $\xi_1^1, \ldots, \xi_{N_1}^1, \ldots, \xi_1^L, \ldots, \xi_{N_L}^L$ are also independent. A new within-class covariance matrix of class C_k is then formed below:

$$\tilde{\mathbf{S}}_k = \mathbf{E}_{\xi^k}\left[\sum_{i=1}^{N_k} \frac{1}{N_k}(\mathbf{x}_i^k - \tilde{\mathbf{u}}_k)(\mathbf{x}_i^k - \tilde{\mathbf{u}}_k)^{\mathrm{T}}\right] = \hat{\mathbf{S}}_k + \frac{1}{N_k}\mathbf{\Omega}_{C_k} \tag{9}$$

So a new within-class covariance matrix is established by:

$$\tilde{\mathbf{S}}_w = \sum_{k=1}^{L} \frac{N_k}{N}\tilde{\mathbf{S}}_k = \hat{\mathbf{S}}_w + \frac{1}{N}\sum_{k=1}^{L}\mathbf{\Omega}_{C_k} = \hat{\mathbf{S}}_w + \mathbf{S}_w^{\Delta} \tag{10}$$

where $\mathbf{S}_w^{\Delta} = \frac{1}{N}\sum_{k=1}^{L}\mathbf{\Omega}_{C_k}$. Next, following equalities (2) and (4), we get

$$\begin{aligned}
&\frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{L}\frac{N_k}{N} \times \frac{N_j}{N}(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j)(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j)^{\mathrm{T}} \\
&\quad = \sum_{k=1}^{L}\frac{N_k}{N}(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}})(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}})^{\mathrm{T}},
\end{aligned}$$

where $\tilde{\mathbf{u}} = \sum_{k=1}^{L}\frac{N_k}{N}\tilde{\mathbf{u}}_k = \hat{\mathbf{u}} + \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}\xi_i^k$. Then a new between-class covariance matrix is given by:

$$\begin{aligned}
\tilde{\mathbf{S}}_b &= \mathbf{E}_{\xi}\left[\frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{L}\frac{N_k}{N} \times \frac{N_j}{N}(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j)(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j)^{\mathrm{T}}\right] \\
&= \hat{\mathbf{S}}_b + \mathbf{S}_b^{\Delta}
\end{aligned} \tag{11}$$

where $\mathbf{S}_b^{\Delta} = \sum_{k=1}^{L}\frac{(N-N_k)^2}{N^3}\mathbf{\Omega}_{C_k} + \sum_{k=1}^{L}\frac{N_k}{N^3}\sum_{s=1, s\neq k}^{L}(N_s\mathbf{\Omega}_{C_s})$. The details of the derivation of Eq. (9) and (11) can be found in Appendix A.

From the above analysis, a new formulation of Fisher's LDA called *perturbation LDA* (P-LDA) is given by the following theorem.

Theorem 1. (P-LDA) *Under the Gaussian distribution of within-class data, perturbation LDA (P-LDA) finds a linear projection matrix* $\tilde{\mathbf{W}}_{opt}$ *such that*:

$$\begin{aligned}
\tilde{\mathbf{W}}_{opt} &= arg\max_{\mathbf{W}}\frac{trace(\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{S}}_b\mathbf{W})}{trace(\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{S}}_w\mathbf{W})} \\
&= arg\max_{\mathbf{W}}\frac{trace(\mathbf{W}^{\mathrm{T}}(\hat{\mathbf{S}}_b + \mathbf{S}_b^{\Delta})\mathbf{W})}{trace(\mathbf{W}^{\mathrm{T}}(\hat{\mathbf{S}}_w + \mathbf{S}_w^{\Delta})\mathbf{W})}.
\end{aligned} \tag{12}$$

Here, $\mathbf{S}_b^{\Delta}$ *and* $\mathbf{S}_w^{\Delta}$ *are called between-class perturbation covariance matrix and within-class perturbation covariance matrix, respectively.*

Finally, we further interpret the effects of covariance matrices $\tilde{\mathbf{S}}_w$ and $\tilde{\mathbf{S}}_b$ based on Eq. (12). Suppose $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_\ell)$ in Eq. (12), where $\mathbf{w}_m(\in \Re^n)$ is a feature vector. Then for any $\mathbf{W}$ and random vectors $\xi = \{\xi_i^k\}_{i=1,\ldots,N_k}^{k=1,\ldots,L}$, we define:

$$f_b(\mathbf{W}, \xi) = \frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{L}\frac{N_k}{N} \times \frac{N_j}{N}\sum_{m=1}^{l}(\mathbf{w}_m^{\mathrm{T}}(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j))^2, \tag{13}$$

$$f_w(\mathbf{W}, \xi) = \frac{1}{N}\sum_{k=1}^{L}\sum_{i=1}^{N_k}\sum_{m=1}^{l}(\mathbf{w}_m^T(\mathbf{x}_i^k - \tilde{\mathbf{u}}_k))^2. \tag{14}$$

Noting that $\tilde{\mathbf{u}}_k = \hat{\mathbf{u}}_k + \frac{1}{N_k}\sum_{i=1}^{N_k}\xi_i^k$ is the random mean of class C_k, so $f_b(\mathbf{W}, \xi)$ is the average pairwise distance between random means of different classes and $f_w(\mathbf{W}, \xi)$ is the average distance between any sample and the random mean of its corresponding class in a lower-dimensional space. Define the following model:

$$\tilde{\mathbf{W}}_{opt}(\xi) = arg\max_{\mathbf{W}} f_b(\mathbf{W}, \xi)/f_w(\mathbf{W}, \xi).$$

Given specific estimates $\hat{\xi} = \{\hat{\xi}_i^k\}_{i=1,\ldots,N_k}^{k=1,\ldots,L}$, we then can get a projection $\tilde{\mathbf{W}}_{opt}(\hat{\xi})$. In practice, it would be hard to find the proper estimate $\hat{\xi}_i^k$ that can accurately describe the difference between $\mathbf{x}_i^k$ and its expectation $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$. Rather than accurately estimating such $\hat{\xi}_i^k$, we instead consider finding the projection by maximizing the ratio

[2] In this paper the notation "$\wedge$" is always added overhead to the corresponding random vector to indicate that it is an estimate of that random vector. As analyzed later, $\hat{\xi}_{\mathbf{x}}$ does not need to be estimated directly, but a technique will be introduced to estimate the information about $\hat{\xi}_{\mathbf{x}}$.

between the expectation values of $f_b(\mathbf{W}, \xi)$ and $f_w(\mathbf{W}, \xi)$ with respect to ξ such that the uncertainty is considered to be over the domain of ξ. That is:

$$\tilde{\mathbf{W}}_{opt} = arg \max_{\mathbf{W}} \mathbf{E}_{\xi}[f_b(\mathbf{W}, \xi)]/\mathbf{E}_{\xi}[f_w(\mathbf{W}, \xi)]$$
$$= arg \max_{\mathbf{W}} f_b(\mathbf{W})/f_w(\mathbf{W})$$

It can be verified that

$$f_b(\mathbf{W}) = \mathbf{E}_{\xi}[f_b(\mathbf{W}, \xi)] = trace(\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{S}}_b\mathbf{W}) \tag{15}$$

$$f_w(\mathbf{W}) = \mathbf{E}_{\xi}[f_w(\mathbf{W}, \xi)] = trace(\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{S}}_w\mathbf{W}) \tag{16}$$

So, it is exactly the optimization model formulated in Eq. (12). This gives an more intuitive understanding of the effects of covariance matrices $\tilde{\mathbf{S}}_w$ and $\tilde{\mathbf{S}}_b$. Though in P-LDA $\hat{\mathbf{S}}_w$ and $\hat{\mathbf{S}}_b$ are perturbated by $\mathbf{S}_w^{\Delta}$ and $\mathbf{S}_b^{\Delta}$**,** respectively, however in Section 5 we will show $\tilde{\mathbf{S}}_w$ and $\tilde{\mathbf{S}}_b$ will converge to the precise within-class and between-class covariance matrices, respectively. This will show the rationality of P-LDA, since the class empirical mean is almost its expectation value when sample size is large enough and then the perturbation effect could be ignored.

2.2. P-LDA under mixture of Gaussian distribution

This section extends Theorem 1 by altering the class distribution from single Gaussian to mixture of Gaussians [3]. Therefore, the probability density function of a sample $\mathbf{x}$ in class C_k is:

$$p(\mathbf{x}|C_k) = \sum_{i=1}^{I_k} P(i|k)\mathbf{N}(\mathbf{x}|\mathbf{u}_k^i, \Xi_k^i), \tag{17}$$

where $\mathbf{u}_k^i$ is the expectation of $\mathbf{x}$ in the ith Gaussian component (GC) $\mathbf{N}(\mathbf{x}|\mathbf{u}_k^i, \Xi_k^i)$ of class C_k, Ξ_k^i is its covariance matrix and $P(i|k)$ is the prior probability of the ith GC of class C_k. Such density function indicates that any sample $\mathbf{x}$ in class C_k mainly distributes in one of the GC. Therefore, Theorem 1 under single Gaussian distribution can be extended to learning perturbation in each GC. To do so, the clusters within each class should be first determined such that data in each cluster are approximately normally distributed. Then those clusters are labeled as subclasses, respectively. Finally P-LDA is used to learn the discriminant information of all those subclasses. It is similar to the idea of Zhu and Martinez [26] who extended classical Fisher's LDA to the mixture of Gaussian distribution case.

In details, suppose there are I_k GCs (clusters) in class C_k and N_k^i out of all N samples are in the ith GC of class C_k. Let C_k^i denote the ith GC of class C_k. If we denote $\mathbf{x}_{i,s}^k$ as the sth sample of C_k^i, $s = 1, \ldots, N_k^i$, then a perturbation random vector $\xi_{i,s}^k$ for $\mathbf{x}_{i,s}^k$ can be modeled, where $\xi_{i,s}^k \sim \mathbf{N}(\mathbf{0}, \mathbf{\Omega}_{C_k^i})$, $\mathbf{\Omega}_{C_k^i} \in \Re^{n\times n}$, so that $\tilde{\mathbf{x}}_{i,s}^k = \mathbf{x}_{i,s}^k + \xi_{i,s}^k$ is a random vector stochastically describes the expectation of subclass C_k^i, i.e., $\mathbf{u}_k^i$. Then P-LDA can be extended to the mixture of Gaussians case by classifying the subclasses $\{C_k^i\}_{i=1,\ldots,I_k}^{k=1,\ldots,L}$. Thus we get the following theorem[3] a straightforward extension of Theorem 1 and the proof is omitted.

Theorem 2. *Under the Gaussian mixture distribution of data within each class, the projection matrix of perturbation LDA (P-LDA), $\tilde{\mathbf{W}}_{opt}''$, can be found as follows*:

$$\tilde{\mathbf{W}}_{opt}'' = arg \max_{\mathbf{W}} \frac{trace(\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{S}}_b''\mathbf{W})}{trace(\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{S}}_w''\mathbf{W})}$$
$$= arg \max_{\mathbf{W}} \frac{trace(\mathbf{W}^{\mathrm{T}}(\hat{\mathbf{S}}_b'' + \mathbf{S}_b''^{\Delta})\mathbf{W})}{trace(\mathbf{W}^{\mathrm{T}}(\hat{\mathbf{S}}_w'' + \mathbf{S}_w''^{\Delta})\mathbf{W})} \tag{18}$$

where

$$\tilde{\mathbf{S}}_b'' = \mathbf{E}_{\xi''}[\tfrac{1}{2}\textstyle\sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=1}^{I_k}\sum_{s=1}^{I_j}\frac{N_k^i}{N} \times \frac{N_j^s}{N}(\tilde{\mathbf{u}}_k^i - \tilde{\mathbf{u}}_j^s)(\tilde{\mathbf{u}}_k^i - \tilde{\mathbf{u}}_j^s)^{\mathrm{T}}] = \hat{\mathbf{S}}_b'' + \mathbf{S}_b''^{\Delta},$$

$$\mathbf{S}_b''^{\Delta} = \textstyle\sum_{k=1}^{L}\sum_{i=1}^{I_k}\frac{(N-N_k^i)^2}{N^3}\mathbf{\Omega}_{C_k^i} + \sum_{k=1}^{L}\sum_{i=1}^{I_k}\frac{N_k^i}{N^3}\sum_{j=1}^{L}\sum_{s=1,(j,s)\neq(k,i)}^{I_j}(N_j^s\mathbf{\Omega}_{C_j^s}),$$

$$\hat{\mathbf{S}}_b'' = \tfrac{1}{2}\textstyle\sum_{k=1}^{L}\sum_{j=1}^{L}\sum_{i=1}^{I_k}\sum_{s=1}^{I_j}\frac{N_k^i}{N} \times \frac{N_j^s}{N}(\hat{\mathbf{u}}_k^i - \hat{\mathbf{u}}_j^s)(\hat{\mathbf{u}}_k^i - \hat{\mathbf{u}}_j^s)^{\mathrm{T}},$$

$$\tilde{\mathbf{S}}_w'' = \textstyle\sum_{k=1}^{L}\sum_{i=1}^{I_k}\frac{N_k^i}{N}\tilde{\mathbf{S}}_k''^i = \hat{\mathbf{S}}_w'' + \mathbf{S}_w''^{\Delta},$$

$$\tilde{\mathbf{S}}_k''^i = \mathbf{E}_{\xi_{k,i}''}[\textstyle\sum_{s=1}^{N_k^i}\frac{1}{N_k^i}(\mathbf{x}_{i,s}^k - \tilde{\mathbf{u}}_k^i)(\mathbf{x}_{i,s}^k - \tilde{\mathbf{u}}_k^i)^{\mathrm{T}}],$$

$$\mathbf{S}_w''^{\Delta} = \tfrac{1}{N}\textstyle\sum_{k=1}^{L}\sum_{i=1}^{I_k}\mathbf{\Omega}_{C_k^i},$$

$$\hat{\mathbf{S}}_w'' = \tfrac{1}{N}\textstyle\sum_{k=1}^{L}\sum_{i=1}^{I_k}\sum_{s=1}^{N_k^i}(\mathbf{x}_{i,s}^k - \hat{\mathbf{u}}_k^i)(\mathbf{x}_{i,s}^k - \hat{\mathbf{u}}_k^i)^{\mathrm{T}},$$

$$\hat{\mathbf{u}}_k^i = \frac{1}{N_k^i}\textstyle\sum_{s=1}^{N_k^i}\mathbf{x}_{i,s}^k, \tilde{\mathbf{u}}_k^i = \hat{\mathbf{u}}_k^i + \frac{1}{N_k^i}\sum_{s=1}^{N_k^i}\xi_{i,s}^k, i = 1, \ldots, I_k, k = 1, \ldots, L,$$

$$\xi_{k,i}'' = \{\xi_{i,1}^k, \ldots, \xi_{i,N_k^i}^k\}, \xi'' = \{\xi_{1,1}'', \ldots, \xi_{1,I_1}'', \ldots, \xi_{L,1}'', \ldots, \xi_{L,I_L}''\}.$$

3. Estimation of perturbation covariance matrices

For implementation of P-LDA, we need to properly estimate two perturbation covariance matrices $\mathbf{S}_b^{\Delta}$ and $\mathbf{S}_w^{\Delta}$. Parameter estimation is challenging, since it is always ill-posed [3,23] due to limited sample size and the curse of high dimensionality. A more robust and tractable way to overcome this problem is to perform some regularized estimation. It is indeed the motivation here. A method will be suggested to implement P-LDA with parameter estimation in an entire PCA subspace without discarding any nonzero principal component. Unlike the covariance matrix estimation on sample data, we will introduce an indirect way for estimation of the covariance matrices of perturbation random vectors, since the observation values of the perturbation random vectors are hard to be found directly.

For derivation, parameter estimation would focus on P-LDA under single Gaussian distribution, and it could be easily generalized to the Gaussian mixture distribution case by Theorem 2. This section

[3] The designs of $\tilde{\mathbf{S}}_b''$ and $\tilde{\mathbf{S}}_w''$ in the criterion are not restricted to the presented forms. The goal here is just to present a way how to generalize the analysis under single Gaussian case.

is divided into two parts. The first part suggests regularized models for estimation of the parameters, and then a method for parameter estimation is presented in the second part.

3.1. Simplified models for regularized estimation

In this paper, we restrict our attention to the data that are not much heteroscedastic, i.e., class covariance matrices are approximately equal[4] (or not differ too much). It is also in line with one of the conditions when Fisher criterion is optimal [3]. Under this condition, we consider the case when perturbation covariance matrices of all classes are approximately equal. Therefore, the perturbation covariance matrices can be replaced by their average, a pooled perturbation covariance matrix defined in Eq. (19). We obtain Lemma 1 with its proof provided in Appendix B.

Lemma 1. *If the covariance matrices of all perturbation random vectors are replaced by their average, i.e., a pooled perturbation covariance matrix as follows*

$$\boldsymbol{\Omega}_{C_1} = \boldsymbol{\Omega}_{C_2} = \cdots = \boldsymbol{\Omega}_{C_L} = \boldsymbol{\Omega}, \tag{19}$$

then $\mathbf{S}_b^{\Delta}$ *and* $\mathbf{S}_w^{\Delta}$ *can be rewritten as*:

$$\mathbf{S}_b^{\Delta} = \frac{L-1}{N}\boldsymbol{\Omega}, \quad \mathbf{S}_w^{\Delta} = \frac{L}{N}\boldsymbol{\Omega}. \tag{20}$$

Note that when class covariance matrices of data do not differ too much, utilizing pooled covariance matrix to replace individual covariance matrix has been widely used and experimentally suggested to attenuate the ill-posed estimation in many existing algorithms [1,23,24,27–30].

To develop a more simplified model in the entire principal component space, we perform principal component analysis [31] in $\mathbf{X}$ without discarding any nonzero principal component. In practice, the principal components can be acquired from the eigenvectors of the total-class covariance matrix $\hat{\mathbf{S}}_t (=\hat{\mathbf{S}}_w + \hat{\mathbf{S}}_b)$. When the data dimension is much larger than the total sample size, the rank of $\hat{\mathbf{S}}_t$ is at most $N-1$ [5,32], i.e., $\mathrm{rank}(\hat{\mathbf{S}}_t) \leqslant N-1$. In general, $\mathrm{rank}(\hat{\mathbf{S}}_t)$ is always equal to $N-1$. For convenience of analysis, we assume $\mathrm{rank}(\hat{\mathbf{S}}_t) \approx N-1$. It also implies that no information is lost for Fisher's LDA, since all positive principal components are retained [33].

Suppose given the decorrelated data space $\mathbf{X}$, the entire PCA space of dimension $n = N-1$. Based on Eq. (6) and Lemma 1, for any given input sample $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbf{X}$, its corresponding perturbation random vector is $\boldsymbol{\xi}_{\mathbf{x}} = (\xi_{\mathbf{x}}^1, \ldots, \xi_{\mathbf{x}}^n)^{\mathrm{T}} \in \Re^n$, where $\boldsymbol{\xi}_{\mathbf{x}} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega})$. Since $\mathbf{X}$ is decorrelated, the coefficients $x_1, \ldots, x_n$ are approximately uncorrelated. Note that the perturbation variables $\xi_{\mathbf{x}}^1, \ldots, \xi_{\mathbf{x}}^n$ are apparently only correlated to their corresponding uncorrelated coefficients $x_1, \ldots, x_\mathrm{n}$, respectively. Therefore it is able to model $\boldsymbol{\Omega}$ by assuming these random variables $\xi_{\mathbf{x}}^1, \ldots, \xi_{\mathbf{x}}^n$ are uncorrelated each other.[5] Based on this principle, $\boldsymbol{\Omega}$ can be modeled by

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}, \quad \boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2), \tag{21}$$

where σ_i^2 is the variance of $\xi_{\mathbf{x}}^i$. Furthermore, if the average variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2$ is used to replace each individual variance σ_i^2, $i = 1, \ldots, n$, a special model is then acquired by

$$\boldsymbol{\Omega} = \sigma^2\mathbf{I}, \quad \sigma \neq 0, \quad \mathbf{I} \text{ is the } n \times n \text{ identity matrix.} \tag{22}$$

From the statistical point of view, the above simplified models could be interpreted as regularized estimations [25] of $\boldsymbol{\Omega}$ on the perturbation random vectors. It is known that when the dimensionality of data is high, the estimation would become ill-posed (poorly posed) if the number of parameters to be estimated is larger than (comparable to) the number of samples [3,23]. Moreover, estimation of $\boldsymbol{\Omega}$ relates to the information of some expectation value, which, however, is hard to be specified in practice. Hence, regularized estimation of $\boldsymbol{\Omega}$ would be preferred to alleviate the ill-posed problem and obtain a stable estimate in applications. To this end, estimation based on Eq. (22) may be more stable than estimating $\boldsymbol{\Lambda}$, since Eq. (22) can apparently reduce the number of estimated parameters. This would be demonstrated and justified by synthetic data in the experiment.

Finally, this simplified perturbation model is still in line with the perturbation LDA model, since the perturbation matrices $\boldsymbol{\Omega}_{C_k}$ as well as their average $\boldsymbol{\Omega}$ need not to be the accurate expected class covariance matrices but only need to follow the perturbation model given below Eq. (5).

3.2. Estimating parameters

An important issue left is to estimate the variance parameters $\sigma_1^2, \ldots, \sigma_n^2$ and σ^2. The idea is straightforward that the parameters are learned from the generated observation values of perturbation random vectors using maximum likelihood. However, an indirect way is desirable, since it is impossible to find the realizations of perturbation random vectors directly. Hence, our idea turns to find some *sums of perturbation random vectors* based on the perturbation model and then generate their realizations for estimation.

3.2.1. Inferring the sum of perturbation random vectors

Suppose N_k, the number of training samples for class C_k, is larger than 1. Define the average of observed samples in class C_k by excluding $\mathbf{x}_j^k$ as

$$\hat{\mathbf{u}}_k^{-j} = \frac{1}{N_k - 1} \sum_{i=1, i \neq j}^{N_k} \mathbf{x}_i^k, \quad j = 1, \ldots, N_k. \tag{23}$$

It is actually feasible to treat $\hat{\mathbf{u}}_k^{-j}$ as another empirical mean of class C_k. Then, another random mean of class C_k is able to be formulated by:

$$\tilde{\mathbf{u}}_k^{-j} = \frac{1}{N_k - 1} \sum_{i=1, i \neq j}^{N_k} \tilde{\mathbf{x}}_i^k = \hat{\mathbf{u}}_k^{-j} + \frac{1}{N_k - 1} \sum_{i=1, i \neq j}^{N_k} \boldsymbol{\xi}_i^k. \tag{24}$$

Comparing with $\tilde{\mathbf{u}}_k$ the random mean of class C_k in terms of Eq. (8), based on the perturbation model, we know $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{u}}_k^{-j}$ can both stochastically approximate to $\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ by the following specific estimates, respectively:

$$\hat{\hat{\mathbf{u}}}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \hat{\hat{\mathbf{x}}}_i^k = \mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}'], \tag{25}$$

$$\hat{\hat{\mathbf{u}}}_k^{-j} = \frac{1}{N_k - 1} \sum_{i=1, i \neq j}^{N_k} \hat{\hat{\mathbf{x}}}_i^k = \mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}'], \tag{26}$$

where $\hat{\hat{\mathbf{x}}}_i^k = \mathbf{x}_i^k + \hat{\boldsymbol{\xi}}_i^k$, $\hat{\boldsymbol{\xi}}_i^k$ is an estimate of $\boldsymbol{\xi}_i^k$ such that $\mathbf{x}_i^k + \hat{\boldsymbol{\xi}}_i^k = \mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}']$ based on the perturbation model. Hence, we can have the

[4] Discussing variants of Fisher's LDA under unequal class covariance matrices is not the scope of this paper. It is another research topic [39].

[5] It might be in theory a suboptimal strategy. However this assumption is practically useful and reasonable to alleviate the ill-posed estimation problem for high-dimensional data by reducing the number of estimated parameters. In Appendix-D, we show its practical rationality by demonstrating an experimental verification for this assumption on face data sets used in the experiment.

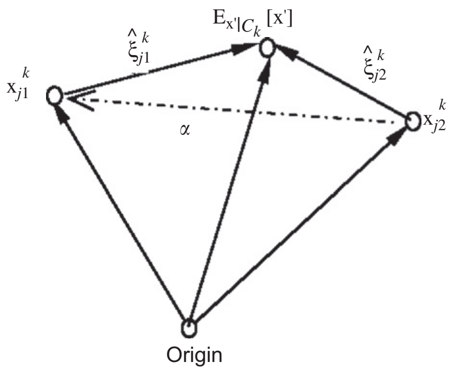

Fig. 1. Geometric interpretation: $\alpha=\mathbf{x}_{j_1}^k-\mathbf{x}_{j_2}^k=\hat{\xi}_{j_2}^k-\hat{\xi}_{j_1}^k$.

relation below:

$$\hat{\mathbf{u}}_k=\hat{\mathbf{u}}_k^{-j}. \tag{27}$$

A geometric interpretation of Eq. (27) can be provided by Fig. 1. Note that $\hat{\mathbf{u}}_k=\hat{\mathbf{u}}_k^{-j_1}=\hat{\mathbf{u}}_k^{-j_2}, j_1\neq j_2$. It therefore yields $\mathbf{x}_{j_1}^k-\mathbf{x}_{j_2}^k=\hat{\xi}_{j_2}^k-\hat{\xi}_{j_1}^k$. According to Eq. (7), this is obviously true because $\hat{\hat{\mathbf{x}}}_i^k=\mathbf{x}_i^k+\hat{\xi}_i^k=\mathbf{E}_{\mathbf{x}'|C_k}[\mathbf{x}'], i=1,\ldots,N_k$.

Now return back to the methodology. Based on Eq. (27) we then have

$$\frac{1}{N_k(N_k-1)}\sum_{i=1,i\neq j}^{N_k}\hat{\xi}_i^k-\frac{1}{N_k}\hat{\xi}_j^k=\hat{\mathbf{u}}_k-\hat{\mathbf{u}}_k^{-j}. \tag{28}$$

Define a new random vector as:

$$\xi_j^{-k}=\frac{1}{N_k(N_k-1)}\left(\sum_{i=1,i\neq j}^{N_k}\xi_i^k\right)-\frac{1}{N_k}\xi_j^k. \tag{29}$$

Based on Lemma 1, we know that the pooled perturbation covariance matrix to be estimated for all $\{\xi_j^k\}$ is $\mathbf{\Omega}$. It is therefore easy to verify the following result:

$$\xi_j^{-k}\sim\mathbf{N}\left(\mathbf{0},\frac{1}{N_k(N_k-1)}\mathbf{\Omega}\right). \tag{30}$$

Actually ξ_j^{-k} is just the *sum of perturbation random vectors* we aim to find. Moreover, Eq. (28) could provide an estimate of ξ_j^{-k} by:

$$\hat{\xi}_j^{-k}=\hat{\mathbf{u}}_k-\hat{\mathbf{u}}_k^{-j}. \tag{31}$$

It therefore avoids the difficulty in finding the observation values $\hat{\xi}_i^k$ directly. Moreover it is known that $\{\hat{\xi}_j^{-k}\}_{j=1,\ldots,N_k}$ follow the same distribution within class C_k, i.e., $\mathbf{N}(\mathbf{0},\frac{1}{N_k(N_k-1)}\mathbf{\Omega})$, so it is feasible to generate N_k observation values $\{\hat{\xi}_1^{-k},\hat{\xi}_2^{-k},\ldots,\hat{\xi}_{N_k}^{-k}\}$ from this distribution. In fact, the empirical mean of the observation values coincides with their expectation with respect to the distribution because of the following equality:

$$\sum_{j=1}^{N_k}\hat{\xi}_j^{-k}=\sum_{j=1}^{N_k}(\hat{\mathbf{u}}_k-\hat{\mathbf{u}}_k^{-j})=\mathbf{0}. \tag{32}$$

3.2.2. *Inferring estimates of $\sigma_1^2,\ldots,\sigma_n^2$ and σ^2*

The estimates of $\sigma_1^2,\ldots,\sigma_n^2$ and σ^2 are given below based on Eq. (30) and the generated $\{\hat{\xi}_j^{-k}\}_{j=1,\ldots,N_k}^{k=1,\ldots,L}$. First we denote

$$\hat{\mathbf{u}}_k^{\Delta_j}=\hat{\mathbf{u}}_k-\hat{\mathbf{u}}_k^{-j}=(\hat{\mathbf{u}}_k^{\Delta_j}(1),\ldots,\hat{\mathbf{u}}_k^{\Delta_j}(n))^{\mathrm{T}}. \tag{33}$$

Then we define $\hat{\sigma}^2(k,j)$ satisfying

$$\frac{1}{N_k(N_k-1)}\hat{\sigma}_i^2(k,j)=(\hat{\mathbf{u}}_k^{\Delta_j}(i))^2. \tag{34}$$

In the uncorrelated space, $\mathbf{\Omega}$ is modeled by $\mathbf{\Omega}=\mathbf{\Lambda}=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$ for approximation, so $\sigma_1^2,\ldots,\sigma_n^2$ are estimated as $\hat{\sigma}_1^2,\ldots,\hat{\sigma}_n^2$ by using maximum likelihood as follows:

$$\hat{\sigma}_i^2=\frac{1}{N}\sum_{k=1}^{L}\sum_{j=1}^{N_k}\hat{\sigma}_i^2(k,j),\quad i=1,\ldots,n. \tag{35}$$

As suggested by Eq. (22), an average variance of $\sigma_1^2,\ldots,\sigma_n^2$ is used, so the estimate $\hat{\sigma}^2$ of σ^2 is obtained below:

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\hat{\sigma}_i^2. \tag{36}$$

Extensive experiments in Section 4 will justify this estimation.

4. Experimental results

The proposed P-LDA algorithm will be evaluated by both synthetic data and face image data. Face images are typical biometric data. Always, the number of available face training samples for each class is very small while the data dimensionality is very high.

This section is divided into three parts. The first and second parts report the experiment results on synthetic data and face data, respectively. In the third part, we verify our parameter estimation strategy on high-dimensional face image data. Through the experiments, two popular classifiers, namely nearest class mean classifier (NCMC) and nearest neighbor classifier (NNC) are selected to evaluate the algorithms. These two kinds of classifiers have been widely used for Fisher's LDA in existing publications. All programs are implemented using Matlab and run on PC with Intel Pentium (R) D CPU 3.40 GHz processor.

4.1. *Synthetic data*

This section is to justify the performances of the proposed P-LDA under Theorems 1 and 2, and show the effects of Eqs. (21) and (22) in modeling P-LDA. Three types of synthetic data following single Gaussian and mixture of Gaussian distributions in each class, respectively are generated in a three-dimensional space. As shown in Tables 1 and 2, for single Gaussian distribution, we consider two cases, in which the covariance matrices are (i) identity covariance matrices multiplied by a constant 0.25 and (ii) equal diagonal covariance matrices, respectively. For each class, 100 samples are generated. For mixture of Gaussian distribution, each class consists of three GC with

Table 1
Overview of the synthetic data (single Gaussian distribution)

Class Id	Mean	Covariance matrix I	Covariance matrix II
Class 1	$(-0.3,-0.5,1.2)^{\mathrm{T}}$	$\begin{pmatrix}0.25&0&0\\0&0.25&0\\0&0&0.25\end{pmatrix}$	$\begin{pmatrix}0.2192&0&0\\0&0.0027&0\\0&0&0.0308\end{pmatrix}$
Class 2	$(-0.1,1.2,1.5)^{\mathrm{T}}$		
Class 3	$(0.9,-0.7,1.1)^{\mathrm{T}}$		

Table 2
Overview of the synthetic data (Gaussian mixture distribution)

Class Id	Mean of first GC	Mean of second GC	Mean of third GC	Covariance matrix
Class 1	$(1,-0.5,-1)^T$	$(0.2,1,0.6)^T$	$(-0.3,-0.5,1.2)^T$	$\begin{pmatrix} 0.0298 & 0 & 0 \\ 0 & 0.6593 & 0 \\ 0 & 0 & 0.5527 \end{pmatrix}$
Class 2	$(-1,-0.5,-1)^T$	$(-0.1,1.2,1.5)^T$	$(1,-1.9,2)^T$	
Class 3	$(0.9,-0.7,1.1)^T$	$(-1.5,0.6,-0.6)^T$	$(1,1.5,1.2)^T$	

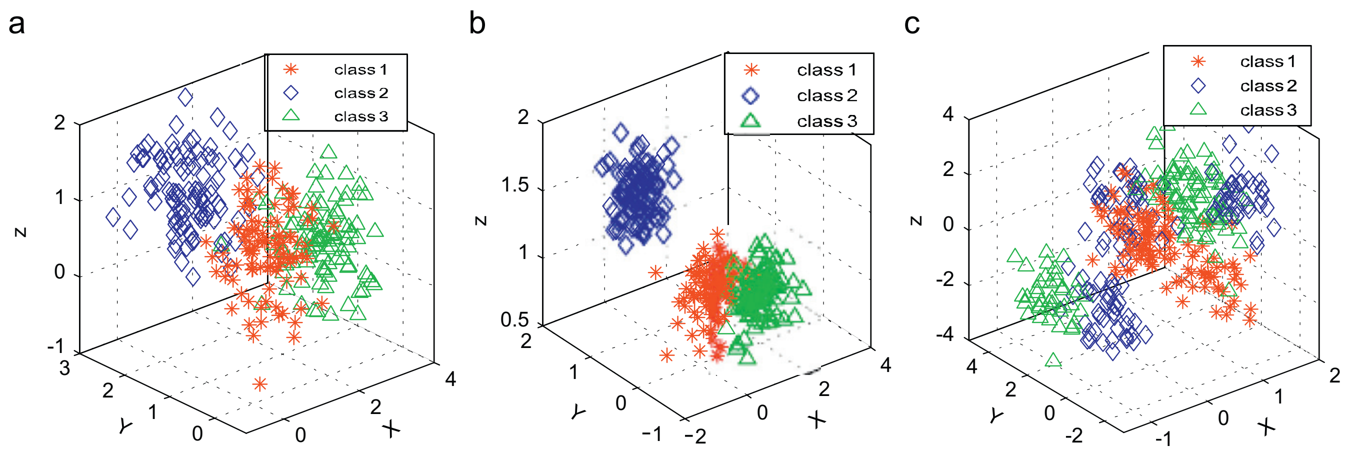

Fig. 2. Illustration of synthetic data: (a) is with equal identity covariance matrices multiplied by 0.25, (b) is with equal diagonal covariance matrices and (c) is with Gaussian mixture distribution.

Table 3
Average accuracy results (equal identity covariance matrices)

Method	Classifier: NCMC			Classifier: NNC		
	$p = 2$ (%)	$p = 5$ (%)	$p = 10$ (%)	$p = 2$ (%)	$p = 5$ (%)	$p = 10$ (%)
P-LDA, Eq. (22)	86.735	90	92.556	85.884	88.772	88.741
P-LDA, Eq. (21)	85.408	90	92.481	83.81	88.491	88.519
Classical Fisher's LDA	82.721	89.439	92.519	81.19	88.281	88.148

Table 4
Average accuracy results (equal diagonal covariance matrices)

Method	Classifier: NCMC			Classifier: NNC		
	$p = 2$ (%)	$p = 5$ (%)	$p = 10$ (%)	$p = 2$ (%)	$p = 5$ (%)	$p = 10$ (%)
P-LDA, Eq. (22)	90.51	93.404	93.481	91.19	93.439	95.296
P-LDA, Eq. (21)	88.469	93.123	93.444	89.354	92.912	95.37
Classical Fisher's LDA	86.803	93.158	93.444	87.993	92.947	95.259

Table 5
Average accuracy results (Gaussian mixture distribution)

Method	Classifier: NCMC				Classifier: NNC			
	$p = 6$ (2) (%)	$p = 9$ (3) (%)	$p = 18$ (6) (%)	$p = 60$ (20) (%)	$p = 6$ (2) (%)	$p = 9$ (3) (%)	$p = 18$ (6) (%)	$p = 60$ (20) (%)
P-LDA (GMM), Eq. (22)	71.257	75.586	77.712	78.556	71.082	72.913	78.725	81.167
P-LDA (GMM), Eq. (21)	68.275	73.874	76.667	78.333	68.363	71.502	78.007	81
Classical Fisher's LDA (GMM)	67.924	73.784	76.601	78.333	68.216	71.291	78.007	81

equal covariance matrices. For each GC, there are 40 samples randomly generated and there are 120 samples for each class. Information about the synthetic data is tabulated in Tables 1 and 2, and the data distributions are illustrated in Fig. 2.

In Tables 3–5, the accuracies with respect to different numbers of training samples for each class are shown, where p indicates the number of training samples for each class. In the mixture of Gaussian distribution case, the bracketed number is the number of training samples from one GC of each class (e.g. "$p = 9$ (3)" means every three samples out of nine training samples of each class are from one of its GCs). For each synthetic data set, we repeat the experiments ten times and the average accuracies are obtained. Since finding GC is not our focus, we assume that those GCs are known for implementation of P-LDA based on Theorem 2. In addition, "P-LDA (GMM), Eq. (22)" means P-LDA is implemented under Gaussian mixture model (GMM) based on Theorem 2 with parameter estimated by Eq. (22); "LDA (GMM)" means classical Fisher's LDA is implemented using a similar scheme to Eq. (18) without the perturbation factors. Note that no singular problem in Fisher's LDA happens in the experiment on synthetic data.

In the single Gaussian distribution case, we find that P-LDA using Eq. (22) outperforms P-LDA using Eq. (21) and classical Fisher's LDA, especially when only two samples for each class are used for training. When the number of training samples for each class increases,

Fig. 3. Some images from the subset of FERET.

Fig. 4. Some images of one subject from the subset of CMU PIE.

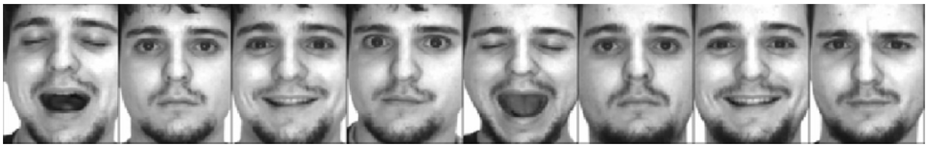

Fig. 5. Images of one subject from the subset of AR.

P-LDA will converge to classical Fisher's LDA, as the class means will be more accurately estimated when more samples are available. In Section 5.1, theoretical analysis would confirm this scenario. Similar results are obtained in the mixture of Gaussian case. These results show that when the number of training samples is small, P-LDA using Eq. (22) can give a more stable and better estimate of the parameter and therefore provide better results.

Table 6
Average recognition accuracy on subset of FERET ($p = 3$)

Method	Classifier: NCMC (%)	Classifier: NNC (%)
P-LDA	87.06	89.29
R-LDA (CV) [13]	86.43	87.96
N-LDA [8]	83.49	83.49
Direct LDA [14]	80.71	78.98
Fisherface [5]	77.25	71.22

Table 7
Average recognition accuracy on subset of CMU PIE

Method	Classifier: NCMC		Classifier: NNC	
	$p = 5$ (%)	$p = 10$ (%)	$p = 5$ (%)	$p = 10$ (%)
P-LDA	78.98	89.94	81.82	93.26
R-LDA (CV) [13]	78.44	89.91	80.43	93.29
N-LDA [8]	74.45	84.98	74.45	84.98
Direct LDA [14]	73.68	85.88	72.73	88.12
Fisherface [5]	72.99	85.49	67.26	82.17

Table 8
Average recognition accuracy on subset of AR

Method	Classifier: NCMC		Classifier: NNC	
	$p = 3$ (%)	$p = 6$ (%)	$p = 3$ (%)	$p = 6$ (%)
P-LDA	92.34	98.28	93.13	98.91
R-LDA (CV) [13]	92.40	98.32	92.81	98.74
N-LDA [8]	91.36	96.43	91.36	96.43
Direct LDA [14]	88.77	97.14	88.42	97.65
Fisherface [5]	86.57	94.66	85.50	94.50

4.2. Face image data

Fisher's LDA based algorithms are popularly used for dimension reduction of high-dimensional data, especially the face images in biometric learning. In this section, the proposed method is applied to face recognition. Since face images are of high dimensionality and only limited samples are available for each person, we implement P-LDA based on Theorem 1 and Eq. (22) with its parameter estimated by Eq. (36).

Three popular face databases, namely FERET [34] database, CMU PIE [35] database and AR database [32], are selected for evaluation. For FERET, a subset consists of 255 persons with four faces for each individual is established. All images are extracted from four different sets, namely Fa, Fb, Fc and the duplicate. Face images in this FERET subset are undergoing illumination variation, age variation and some slight expression variation. For CMU PIE, a subset is established by selecting face images under all illumination conditions with flash in door [35] from the frontal pose, 1/4 left/right profile and below/above in frontal view. There are totally 7140 images and 105 face images for each person in this subset. For AR database, a subset is established by selecting 119 persons, where there are eight images for each person. Face images in this subset are undergoing notable expression variations. All face images are aligned according to their coordinates of the eyes and face centers, respectively. Each image is linearly stretched to the full range of [0,1] and its size is simply normalized to 40×50. Some images are illustrated in Figs. 3–5.

In order to evaluate the proposed model, P-LDA is compared with some Fisher's LDA-based methods including Fisherface [5], nullspace LDA (N-LDA) [8], Direct LDA [14] and regularized LDA with CV CR-LDA (CV) [13], which are popular used for solving the small sample size problem in Fisher's LDA for face recognition.

On each data set, the experiments are repeated 10 times. For each time, p images for each person are randomly selected for training and the rest are for testing. In the tables, the value of p is indicated. Finally, the average recognition accuracies are obtained.

The results are tabulated in Tables 6–8. We see that P-LDA achieves at least 6% and 3% improvements over Direct LDA and

Table 9
Expense of R-LDA (CV)

Method	FERET, $p=3$	CMU PIE, $p=5$	CMU PIE, $p=10$	AR, $p=3$	AR, $p=6$
Time/run (NNC/NCMC)	19∼20 hours	∼1 hours	∼7.5 hours	∼1.2 hours	8.5∼9 hours

Table 10
Average recognition accuracy of P-LDA on FERET data set: "P-LDA with manually selected optimal parameter" vs. "P-LDA with parameter estimation"

Method	Classifier: NCMC			Classifier: NNC		
	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)
P-LDA with manually selected optimal parameter	87.25	90.16	91.80	89.33	91.29	92.12
P-LDA with parameter estimation	87.06	90.35	91.88	89.29	91.25	92.08

Table 11
Average recognition accuracy of P-LDA on CMU PIE data set: "P-LDA with manually selected optimal parameter" vs. "P-LDA with parameter estimation"

Method	Classifier: NCMC			Classifier: NNC		
	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)
P-LDA with manually selected optimal parameter	79.02	83.93	86.44	81.95	85.45	87.33
P-LDA with parameter estimation	78.98	83.89	86.40	81.82	85.12	86.97

N-LDA, respectively, on FERET database, and achieves more than 4% improvement over Fisherface, Direct LDA and N-LDA on CMU PIE database. On AR subset, P-LDA also gets significant improvements over Fisherface and Direct LDA and gets more than 1% improvement over N-LDA. Note that no matter using NNC or NCMC, the results of N-LDA are the same, because N-LDA will map all training samples of the same class into the corresponding class empirical mean in the reduce space [7].

In addition, a related method R-LDA with CV parameter[6] is also conducted for comparison. On FERET, P-LDA gets more than one percent improvement when using NNC and gets about 0.6% improvement when using NCMC. On CMU, when $p=5$, P-LDA gets 1.4% improvement over R-LDA using NNC and 0.5% improvement using NCMC; when $p=10$, P-LDA and R-LDA gets almost the same performances. On AR subset, the performances of P-LDA and R-LDA are also similar. Though R-LDA gets similar performance to P-LDA in some cases, however, as reported in Table 9, R-LDA is extremely computationally expensive due to the CV process. In our experiments, P-LDA can finish in much less than one minute for each run, while R-LDA using CV technique takes more than one hour. More comparison between P-LDA and R-LDA could be found in Section 5.2. It will be analyzed later that R-LDA can be seen as a semi-perturbation LDA, which gives a novel understanding to R-LDA. It would also be explored that the proposed perturbation model actually can suggest an effective and efficient way for the regularized parameter estimation in R-LDA. Therefore, P-LDA is much more efficient and still performs better.

Although Fisherface, Direct LDA, N-LDA and R-LDA are also proposed for extraction of discriminant features in the undersampled case, they mainly address the singularity problem of the within-class matrix, while P-LDA addresses the perturbation problem in Fisher criterion due to the difference between a class empirical mean and its expectation value. Noting that P-LDA using model (21) and (22) can also solve the singularity problem, this suggests alleviating the perturbation problem is useful to further enhance the Fisher criterion.

In addition, the above results as well as the results on synthetic data sets also indicate that when the number of training samples is large, the differences between P-LDA and the compared LDA based algorithms become small. This is true according to the perturbation analysis given in this paper, since the estimates of the class means will be more accurate when training samples for each class become more sufficient. Noting also that the difference between P-LDA and R-LDA is small when p is large on CMU and AR, it implies the impact of the perturbation model in estimation of the between-class covariance information will become minor as the number of training samples increases. In Section 5.1, we would give more theoretical analysis.

4.3. Parameter verification

In the last two subsections, we show that P-LDA using Eq. (22) gives good results on both synthetic and face image data, particularly when the number of training samples is small. In this section, we will have extensive statistics of the performances of P-LDA on FERET and CMU PIE if the parameter σ^2 is set to be other values. We compare the proposed P-LDA with parameter estimation with the best scenario selected manually.

The detailed procedure of the experiments is listed as follows.

Step (1): Prior values of σ^2 are extensively sampled. We let $\sigma^2 = \frac{\eta}{1-\eta}$, $0<\eta<1$, so that $\sigma^2 \in (0,+\infty)$. Then 1999 points are sampled for η between 0.0005 and 0.9995 with interval 0.0005. Finally, 1999 sampled values of σ^2 are obtained.

Step (2): Evaluate the performance of P-LDA with respect to each sampled value of σ^2. We call each P-LDA with respect to a sampled value of σ^2 a *model*.

Step (3): We compare the P-LDA model with parameter σ^2 estimated by the methodology suggested in Section 3.2 against the best one among all models of P-LDA got at step (2).

The average recognition rate of each model of P-LDA is obtained by using the same procedure run on FERET and CMU PIE databases. We consider the case when p, the number of training samples for each class, is equal to three on FERET and equal to five on CMU. For clear description, the P-LDA model with parameter estimated using the methodology suggested in Section 3.2 is called "*P-LDA with parameter estimation*", whereas we call the P-LDA model with

[6] On FERET, three-fold CV is performed; On CMU, five-fold CV is performed when $p=5$ and 10-fold CV is performed when $p=10$; On AR, three-fold CV is performed when $p=3$ and six-fold CV is performed when $p=6$. The candidates of the regularization parameter λ are sampled from 0.005 to 1 with step 0.005. In the experiment, the three-fold CV is repeated ten times on FERET. On CMU, the five-fold and 10-fold CV are repeated six and three times, respectively; on AR, the three-fold and six-fold CV are repeated 10 and 5 times, respectively. So, each CV parameter is determined via its corresponding 30 round CV classification.

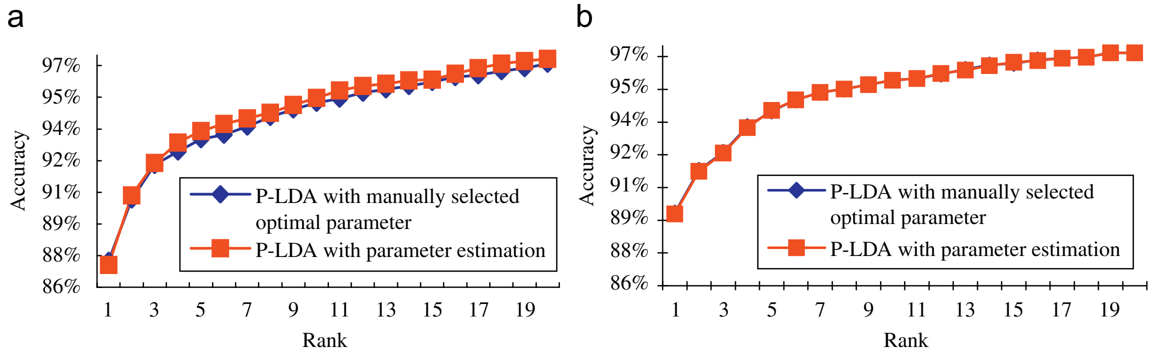

Fig. 6. "P-LDA with manually selected optimal parameter" vs. "P-LDA with parameter estimation" on FERET.

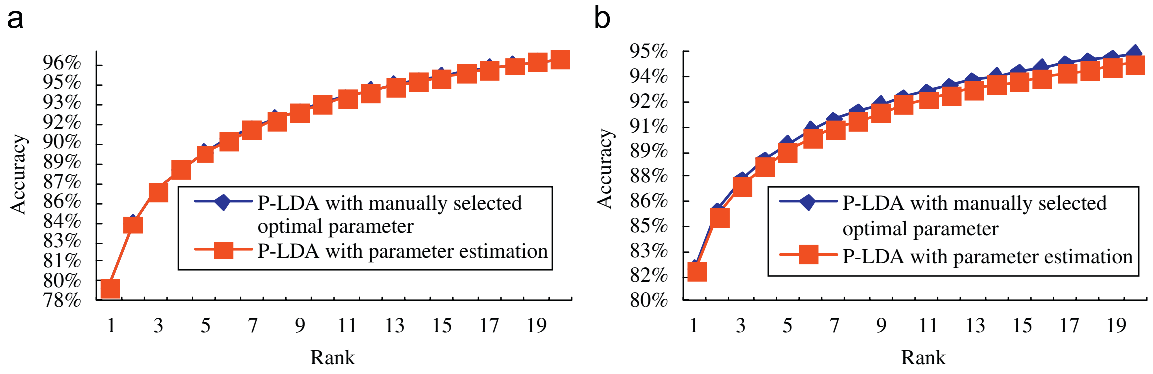

Fig. 7. "P-LDA with manually selected optimal parameter" vs. "P-LDA with parameter estimation" on CMU.

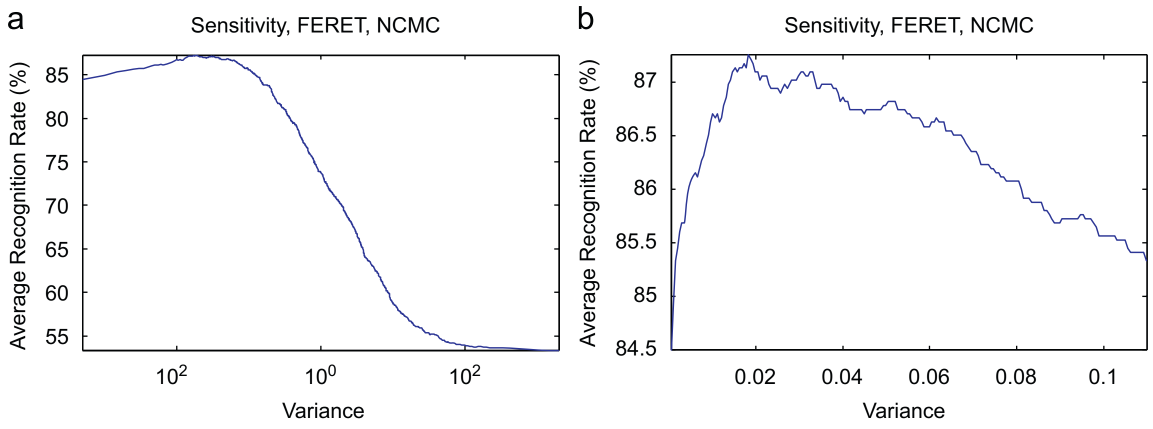

Fig. 8. Classifier: NCMC. (a) The performance of P-LDA as a function of σ^2 (x-axis) on FERET, where the horizontal axis is scaled logarithmically and (b) the enlarged part of (a) near the peak of the curve where σ^2 is small.

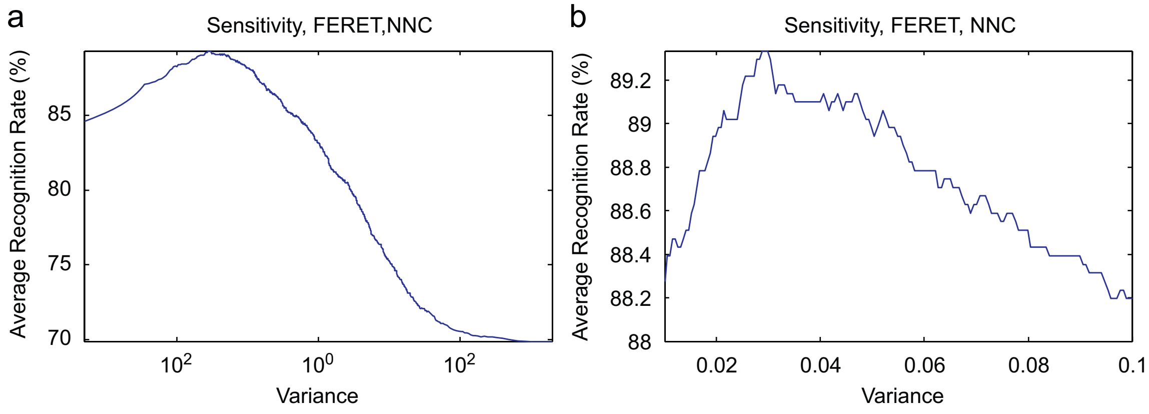

Fig. 9. Classifier: NNC. (a) The performance of P-LDA as a function of σ^2 (x-axis) on FERET, where the horizontal axis is scaled logarithmically; (b) the enlarged part of (a) near the peak of the curve where σ^2 is small.

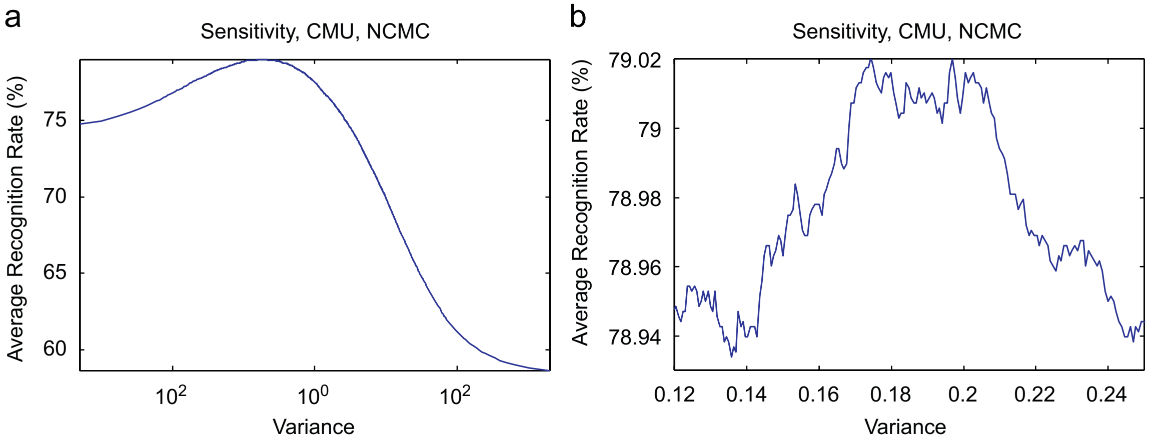

Fig. 10. Classifier: NCMC. (a) The performance of P-LDA as a function of σ^2 (x-axis) on CMU PIE, where the horizontal axis is scaled logarithmically and (b) the enlarged part of (a) near the peak of the curve where σ^2 is small.

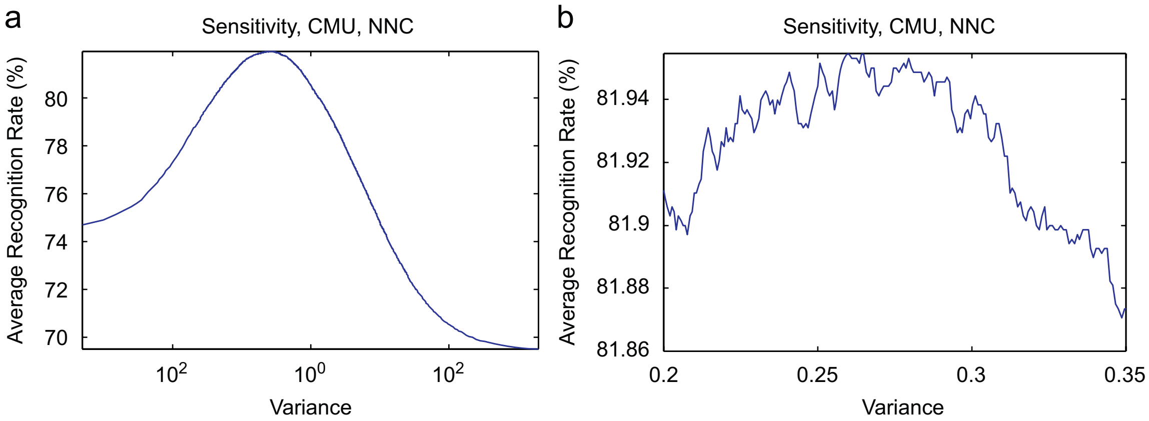

Fig. 11. Classifier: NNC. (a) The performance of P-LDA as a function of σ^2 (x-axis) on CMU PIE, where the horizontal axis is scaled logarithmically; (b) the enlarged part of (a) near the peak of the curve where σ^2 is small.

respect to the best σ^2 selected from the 1999 sampled values "*P-LDA with manually selected optimal parameter*". Comparison results of the rank 1–3 accuracies are reported in Tables 10 and 11. Figs. 6 and 7 show the ranking accuracies of these two models. It shows that the difference of rank 1 accuracies between two models is less than 0.2% in general.

To evaluate the sensitivity of P-LDA on σ^2, the performance of P-LDA as a function of σ^2 is shown from Fig. 8 to Fig. 9 using NCMC and NNC classifiers, respectively. The overall sensitivity of P-LDA on σ^2 for FERET data set is described in Fig. 8(a), where the horizontal axis is on a logarithmic scale. Fig. 8(b) shows the enlarged part of Fig. 8(a) near the peak of the curve where σ^2 is small. Similarly, Figs. 10 and 11 show the result on CMU PIE. They show it may be hard to obtain an optimal estimate of σ^2, but interestingly it is shown in Tables 10 and 11 and Figs. 6 and 7 that the suggested methodology in Section 3.2 works well. It is apparent that selecting the best parameter manually using an extensive search would be time consuming, while P-LDA using the proposed methodology for parameter estimation costs much less than one minute. So the suggested methodology is computationally efficient.

5. Discussion

As shown in the experiment, the number of training samples for each class is really an impact of the performance of P-LDA. In this section, we explore some theoretical properties of P-LDA and the convergence of P-LDA will be shown. We also discuss P-LDA with some related methods.

5.1. Admissible condition of P-LDA

Suppose L is fixed. Since the entries of all perturbation covariance matrices are bounded,[7] it is easy to obtain $\mathbf{S}_b^{\Delta}=O(\frac{1}{N})$ and $\mathbf{S}_w^{\Delta}=O(\frac{1}{N})$, i.e., the perturbation factor $\mathbf{S}_b^{\Delta}\rightarrow\mathbf{O}$, $\mathbf{S}_w^{\Delta}\rightarrow\mathbf{O}$ when $\frac{1}{N}\rightarrow 0$, where $\mathbf{O}$ is the zero matrix. Here, for any matrix $\mathbf{A}=\mathbf{A}(\beta)$ of which each nonzero entry depends on β, we say $\mathbf{A}=O(\beta)$ if the degree[8] of $\mathbf{A}\rightarrow\mathbf{O}$ is comparable to the degree of $\beta\rightarrow 0$.

However, if L is a variant, i.e., the increase of the sample size may be partly due to the increase of the amount of classes, then $\mathbf{S}_b^{\Delta}\neq O(\frac{1}{N})$ and $\mathbf{S}_w^{\Delta}\neq O(\frac{1}{N})$. Suppose any covariance matrix $\mathbf{\Omega}_{C_k}$ is lower (upper) bounded by $\mathbf{\Omega}_{lower}$ if and only if $\mathbf{\Omega}_{lower}(i,j)\leqslant\mathbf{\Omega}_{C_k}(i,j)(\mathbf{\Omega}_{C_k}(i,j)\leqslant\mathbf{\Omega}_{upper}(i,j))$ for any (i,j). Then the following lemma gives an essential view, and its proof is given in Appendix C.

Lemma 2. *If all nonzero perturbation covariance matrices* $\mathbf{\Omega}_{C_k}$, $k=1,\ldots,L$, *are lower bounded by* $\mathbf{\Omega}_{lower}$ *and upper bounded by*

[7] We say a matrix is bounded if and only if all entries of this matrix are bounded.

[8] The degree of $\mathbf{A}=\mathbf{A}(\beta)\rightarrow\mathbf{O}$ depending on β is defined to be the smallest degree for $\mathbf{A}(i,j)\rightarrow 0$ depending on β, where $\mathbf{A}(i,j)$ is any nonzero entry of $\mathbf{A}$. For example, $\mathbf{A}=[\beta\ \beta^2]$, then the degree of $\mathbf{A}\rightarrow\mathbf{O}$ is 1 and $\mathbf{A}=O(\beta)$.

Table 12
Average recognition accuracy of R-LDA on FERET data set: "R-LDA with manually selected optimal parameter" vs. "R-LDA using perturbation model" ($p = 3$)

Method	Classifier: NCMC			Classifier: NNC		
	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)
R-LDA with manually selected optimal parameter	86.78	90.24	91.69	88.27	90.16	91.25
R-LDA (CV)	86.43	89.96	91.49	87.96	90.26	91.33
R-LDA using perturbation model	86.47	90.00	91.69	88.08	90.20	91.49

Table 13
Average recognition accuracy of R-LDA on CMU PIE data set: "R-LDA with manually selected optimal parameter" vs. "R-LDA using perturbation model" ($p = 5$)

Method	Classifier: NCMC			Classifier: NNC		
	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)
R-LDA with manually selected optimal parameter	78.60	83.42	85.88	80.50	84.08	85.98
R-LDA (CV)	78.44	83.27	85.72	80.43	84.05	85.94
R-LDA using perturbation model	78.24	83.51	86.13	80.18	84.12	86.14

$\mathbf{\Omega}_{upper}$, *where* $\mathbf{\Omega}_{lower}$ *and* $\mathbf{\Omega}_{upper}$ *are independent of L and N, then it is true that* $\mathbf{S}_b^{\Delta} = O(\frac{L}{N})$ *and* $\mathbf{S}_w^{\Delta} = O(\frac{L}{N})$.

The condition of Lemma 2 is valid in practice, because the data space is always compact and moreover it is always a Euclidean space of finite dimension. In particular, from Eq. (20), it could be found that the perturbation matrices depend on the average sample size for each class. Based on Theorem 1, we finally have the following proposition.

Proposition 1. (**Admissible condition of P-LDA**) *P-LDA depends on the average number of samples for each class. That is* $\mathbf{S}_b^{\Delta} = O(\frac{L}{N})$ *and* $\mathbf{S}_w^{\Delta} = O(\frac{L}{N})$, *i.e.,* $\mathbf{S}_b^{\Delta} \to \mathbf{O}$, $\mathbf{S}_w^{\Delta} \to \mathbf{O}$ *when* $\frac{L}{N} \to 0$.

It is intuitive that some estimated class means are unstable when the average sample size for each class is small.[9] This also shows what P-LDA targets for is different from the singularity problem in Fisher's LDA, which will be solved if the total sample size is large enough. Moreover the experiments on synthetic data in Section 4.1 could provide the support to Proposition 1, as the difference between P-LDA and classical Fisher's LDA become smaller when the average sample size for each class becomes larger.

5.2. Discussion with related approaches

5.2.1. P-LDA vs. R-LDA

Regularized LDA (R-LDA) is always modeled by the following criterion:

$$\mathbf{W}_{opt} = arg \max_{\mathbf{W}} \frac{\mathrm{trace}(\mathbf{W}^{\mathrm{T}}\hat{\mathbf{S}}_b\mathbf{W})}{\mathrm{trace}(\mathbf{W}^{\mathrm{T}}(\hat{\mathbf{S}}_w + \lambda\mathbf{I})\mathbf{W})}, \quad \lambda > 0. \tag{37}$$

Sometimes, a positive diagonal matrix is used to replace $\lambda\mathbf{I}$ in the above equality.

Generally, the formulation of P-LDA in Section 2 is different from the form of R-LDA. Although the formulation of R-LDA looks similar to the simplified model of P-LDA in Section 3, *the motivation and objective are totally different*. Details are discussed as follows.

1. P-LDA is proposed by learning the difference between a class empirical mean and its corresponding expectation value as well as its impact to Fisher criterion, whereas R-LDA is originally proposed for the singularity problem [9,10,13] because $\hat{\mathbf{S}}_w + \lambda\mathbf{I}$ is positive with $\lambda > 0$.
2. In P-LDA, the effects of $\mathbf{S}_b^{\Delta}$ and $\mathbf{S}_w^{\Delta}$ are known based on the perturbation analysis in theory. In contrast, R-LDA still does not clearly tell how $\lambda\mathbf{I}$ has effect on $\hat{\mathbf{S}}_w$ in a pattern recognition sense. Although Zhang et al. [12] presented a connection between the regularization network algorithms and R-LDA from a least square view, it still lacks interpretation how regularization can has effect on within-class and between-class covariance matrices simultaneously and also lacks parameter estimation.
3. P-LDA tells the convergence of perturbation factors by Proposition 1. However, R-LDA does not tell it in theory. The singularity problem R-LDA addresses is in nature an implementation problem and it would be solved when the total sample size is sufficiently large, while it does not imply the average sample size for each class is also sufficiently large in this situation.
4. P-LDA is developed when data of each class follow either single Gaussian distribution or Gaussian mixture distribution, but R-LDA has not considered the effect of data distribution.
5. In P-LDA, scheme for parameter estimation is an intrinsic methodology derived from the perturbation model itself. For R-LDA, a separated algorithm is required, such as the CV method, which is so far popular. However, CV seriously lies on a discrete set of candidate parameters. In general, CV is always time consuming.

Interestingly, if the proposed perturbation model is imposed on R-LDA, i.e., R-LDA is treated as a *semi-perturbation* Fisher's LDA, where only within-class perturbation $\mathbf{S}_w^{\Delta}$ is considered and the factor $\mathbf{S}_b^{\Delta}$ is ignored, then the methodology in Section 3 may provide an interpretation how the term $\lambda\mathbf{I}$ has its effect in the entire PCA space. This novel view to R-LDA can give the advantage in applying the proposed perturbation model for an efficient and effective estimation of the regularized parameter λ in R-LDA. To justify this, similar comparisons on FERET and CMU subsets between "R-LDA with manually selected optimal parameter" and "R-LDA using perturbation model" are performed in Tables 12 and 13, where "R-LDA with manually selected optimal parameter" is implemented similarly to "P-LDA with manually selected optimal parameter" as demonstrated in Section 4.3. For reference, the results of R-LDA (CV) are also shown. We find that "R-LDA using perturbation model" extremely approximates to "R-LDA with manually selected optimal parameter" and achieves

[9] With suitable training samples, the class means may be well estimated, but selection of training samples is beyond the scope of this paper.

almost the same performances as R-LDA (CV). This indicates that the proposed perturbation model could also be an alternative, practical and efficient way for parameter estimation in R-LDA.

5.2.2. Other comparisons

Recently, a related work called median LDA has been proposed by Yang et al. [36], in which they addressed the estimation of the class mean in Fisher's LDA by using median mean. However, the analysis of the perturbation impact of the estimation of class mean on two covariance matrices in Fisher criterion is not systematically and theoretically presented.

Another related work is known as the concentration inequality (learning) in learning theory [37,38], such as Hoeffding's inequality that describes the difference between empirical mean and its expectation. But only statistical bound is reported. The bound may be loose and the effect of such difference has not been integrated into the discriminate learning algorithm such as Fisher's LDA. In contrast, in P-LDA, a *random mean* is modeled to stochastically characterize the expectation value of each class. P-LDA has been developed by integrating the perturbation between the empirical mean of each class and its expectation value into the learning process.

6. Conclusion

This paper addresses a fundamental research issue in Fisher criterion—the class empirical mean is equal to its expectation. This is one of the assumptions made in deriving the Fisher's LDA formulation for practical computation. However, in many pattern recognition applications, especially the biometric learning, this assumption may not be true. In view of this, we introduce perturbation random vectors to learn the effect of the difference between the class empirical mean and its expectation in Fisher criterion, and then a new formulation, namely perturbation LDA (P-LDA) is developed. The perturbation analysis has finally yielded new forms of within-class and between-class covariance matrices by integrating some perturbation factors in Fisher criterion. A complete theory and mathematical derivation of P-LDA under single Gaussian distribution and mixture of Gaussian distribution of data in each class are developed, respectively. For practical implementation of the proposed P-LDA method, a technique for estimation of the covariance matrices of perturbation random vectors is also developed. Moreover, the proposed perturbation model also gives a novel view to regularized LDA (R-LDA), resulting in an efficient and effective estimation of regularized parameter. Experiments have been performed to evaluate P-LDA and do comparison with recently developed popular Fisher's LDA-based algorithms for solving the small sample size problem. The results show that the proposed P-LDA algorithm is efficient and obtains better performances. In future, the perturbation model in Fisher's LDA may be further developed. In this paper, P-LDA relies on Gaussian assumption of data distribution in each class. Though P-LDA under mixture of Gaussians is also developed, it is currently required that the Gaussian components (GC) are first found, which is still an active research issue in pattern recognition. Therefore, non-parametric technique may be considered for its future development.

Acknowledgments

This project was supported by the NSFC (60675016, 60633030), the 973 Program (2006CB303104), NSF of GuangDong (06023194, 2007B030603001) and Earmarked Research Grant HKBU2113/06E from Hong Kong Research Grant Council. The authors would also like to thank the great efforts made by (associate) editor and all reviewers for improvement of this paper.

Appendix A. Derivation of Eqs. (9) and (11)

$$\tilde{\mathbf{S}}_k = \mathbf{E}_{\xi^k}\left[\sum_{i=1}^{N_k}\frac{1}{N_k}(\mathbf{x}_i^k - \tilde{\mathbf{u}}_k)(\mathbf{x}_i^k - \tilde{\mathbf{u}}_k)^{\mathrm{T}}\right]$$

$$= \sum_{i=1}^{N_k}\frac{1}{N_k}(\mathbf{x}_i^k - \hat{\mathbf{u}}_k)(\mathbf{x}_i^k - \hat{\mathbf{u}}_k)^{\mathrm{T}}$$

$$+ \sum_{i=1}^{N_k}\mathbf{E}_{\xi^k}\left[\frac{1}{(N_k)^3}\left(\sum_{j=1}^{N_k}\xi_j^k\right)\left(\sum_{j=1}^{N_k}\xi_j^k\right)^{\mathrm{T}}\right]$$

$$= \hat{\mathbf{S}}_k + \frac{1}{(N_k)^2}\sum_{j=1}^{N_k}\mathbf{E}_{\xi_j^k}\left[\left(\xi_j^k\right)\left(\xi_j^k\right)^{\mathrm{T}}\right]$$

$$= \hat{\mathbf{S}}_k + \frac{1}{N_k}\mathbf{\Omega}_{C_k}.$$

$$\tilde{\mathbf{S}}_b = \mathbf{E}_{\xi}\left[\frac{1}{2}\sum_{k=1}^{L}\sum_{j=1}^{L}\frac{N_k}{N}\times\frac{N_j}{N}(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j)(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}}_j)^{\mathrm{T}}\right]$$

$$= \mathbf{E}_{\xi}\left[\sum_{k=1}^{L}\frac{N_k}{N}(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}})(\tilde{\mathbf{u}}_k - \tilde{\mathbf{u}})^{\mathrm{T}}\right]$$

$$= \sum_{k=1}^{L}\frac{N_k}{N}(\hat{\mathbf{u}}_k - \hat{\mathbf{u}})(\hat{\mathbf{u}}_k - \hat{\mathbf{u}})^{\mathrm{T}}$$

$$+ \sum_{k=1}^{L}\frac{N_k}{N}\left(\mathbf{E}_{\xi}\left[\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\xi_i^k - \frac{1}{N}\sum_{s=1}^{L}\sum_{i=1}^{N_s}\xi_i^s\right)\right.\right.$$

$$\left.\left.\times\left(\frac{1}{N_k}\sum_{i=1}^{N_k}\xi_i^k - \frac{1}{N}\sum_{s=1}^{L}\sum_{i=1}^{N_s}\xi_i^s\right)^{\mathrm{T}}\right]\right)$$

$$= \hat{\mathbf{S}}_b + \sum_{k=1}^{L}\frac{N_k}{N}\mathbf{E}_{\xi}\left[\left(\frac{N-N_k}{N_kN}\sum_{i=1}^{N_k}\xi_i^k - \frac{1}{N}\sum_{s=1,s\neq k}^{L}\sum_{i=1}^{N_s}\xi_i^s\right)\right.$$

$$\left.\times\left(\frac{N-N_k}{N_kN}\sum_{i=1}^{N_k}\xi_i^k - \frac{1}{N}\sum_{s=1,s\neq k}^{L}\sum_{i=1}^{N_s}\xi_i^s\right)^{\mathrm{T}}\right]$$

$$= \hat{\mathbf{S}}_b + \sum_{k=1}^{L}\frac{N_k}{N}\left(\frac{N-N_k}{N_kN}\right)^2\left(\sum_{i=1}^{N_k}\mathbf{E}_{\xi_i^k}[(\xi_i^k)(\xi_i^{k^{\mathrm{T}}})]\right)$$

$$+ \sum_{k=1}^{L}\frac{N_k}{N}\left(\frac{1}{N}\right)^2\left(\sum_{s=1,s\neq k}^{L}\sum_{i=1}^{N_s}\mathbf{E}_{\xi_i^s}[(\xi_i^s)(\xi_i^{s^{\mathrm{T}}})]\right)$$

$$= \hat{\mathbf{S}}_b + \sum_{k=1}^{L}\frac{N_k}{N}\left(\frac{N-N_k}{N_kN}\right)^2 N_k\mathbf{\Omega}_{C_k}$$

$$+ \sum_{k=1}^{L}\frac{N_k}{N}\left(\frac{1}{N}\right)^2\sum_{s=1,s\neq k}^{L}(N_s\mathbf{\Omega}_{C_s})$$

$$= \hat{\mathbf{S}}_b + \sum_{k=1}^{L}\frac{(N-N_k)^2}{N^3}\mathbf{\Omega}_{C_k} + \sum_{k=1}^{L}\frac{N_k}{N^3}\sum_{s=1,s\neq k}^{L}(N_s\mathbf{\Omega}_{C_s})$$

$$= \hat{\mathbf{S}}_b + \mathbf{S}_b^{\Delta}.$$

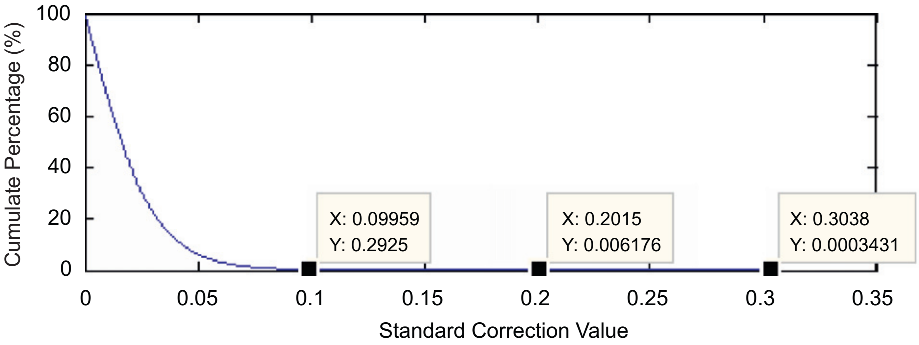

Fig. 12. $F(\beta)$ (y-axis) vs. β (x-axis) on subset of FERET ($p=3$).

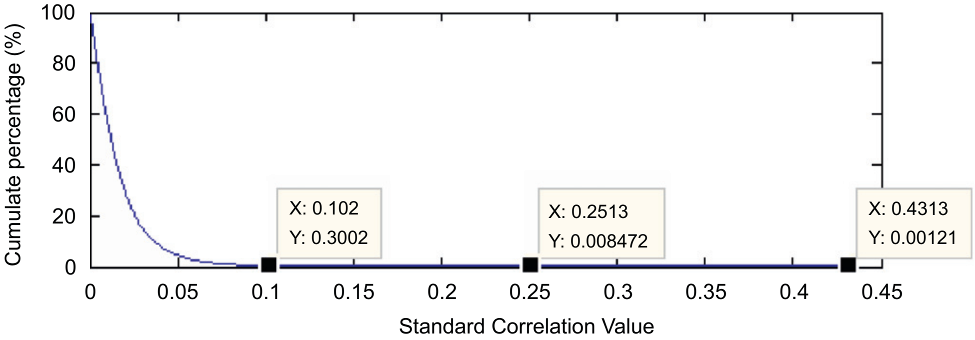

Fig. 13. $F(\beta)$ (y-axis) vs. β (x-axis) on subset of CMU PIE ($p=6$).

Appendix B. Proof of Lemma 1

Proof. $\mathbf{S}_w^{\varDelta}$ is true obviously and the proof is for $\mathbf{S}_b^{\varDelta}$ here. Since $\sum_{s=1,s\neq k}^{L} N_s = N - N_k$, $k=1,\ldots,L$, then:

$$\mathbf{S}_b^{\varDelta} = \sum_{k=1}^{L} \frac{(N-N_k)^2}{N^3}\mathbf{\Omega} + \sum_{k=1}^{L} \frac{N_k}{N^3} \sum_{s=1,s\neq k}^{L} (N_s\mathbf{\Omega})$$
$$= \frac{L-1}{N}\mathbf{\Omega}. \qquad \square$$

Appendix C. Proof of Lemma 2

Proof. For convenience, we denote $\mathbf{\Omega}_{lower} \leqslant \mathbf{\Omega}_{C_k} (\mathbf{\Omega}_{C_k} \leqslant \mathbf{\Omega}_{upper})$ which means $\mathbf{\Omega}_{C_k}$ is lower (upper) bounded by $\mathbf{\Omega}_{lower}(\mathbf{\Omega}_{upper})$. Similarly to the proof in Lemma 1, it is easy to have the following relations:

$$\frac{L-1}{N}\mathbf{\Omega}_{lower} \leqslant \mathbf{S}_b^{\varDelta} \leqslant \frac{L-1}{N}\mathbf{\Omega}_{upper},$$
$$\frac{L}{N}\mathbf{\Omega}_{lower} \leqslant \mathbf{S}_w^{\varDelta} \leqslant \frac{L}{N}\mathbf{\Omega}_{upper}. \tag{C1}$$

Since $\mathbf{\Omega}_{lower}$ and $\mathbf{\Omega}_{upper}$ are independent of L and N and $\frac{L}{N} \to 0$ implies $\frac{1}{N} \to 0$ for $L \geqslant 1$, so it is true that $\mathbf{S}_b^{\varDelta} = O(\frac{L}{N})$ and $\mathbf{S}_w^{\varDelta} = O(\frac{L}{N})$. $\square$

Appendix D. Experimental verification

We here experimentally provide support for the suboptimal but practical strategy used to model $\mathbf{\Omega}$ by assuming random variables $\xi_{\mathbf{x}}^1,\ldots,\xi_{\mathbf{x}}^n$ to be uncorrelated each other in the entire principal component space in Section 3.1. We show that this assumption is really practically useful. Recall the parameter estimation in Section 3.2 where we get $\boldsymbol{\xi}_j^{-k} \sim \mathbf{N}(\mathbf{0}, \frac{1}{N_k(N_k-1)}\mathbf{\Omega})$. Hence a general estimate $\hat{\mathbf{\Omega}}$ for $\mathbf{\Omega}$ is calculated by $\hat{\mathbf{\Omega}} = \frac{1}{N}\sum_{k=1}^{L} N_k(N_k-1)\sum_{j=1}^{N_k}(\hat{\boldsymbol{\xi}}_j^{-k})(\hat{\boldsymbol{\xi}}_j^{-k})^{\mathrm{T}}$ using the generated observation values $\{\hat{\boldsymbol{\xi}}_j^{-k}\}_{j=1,\ldots,N_k}^{k=1,2,\ldots,L}$. Then we can have statistics of the cumulate percentage $F(\beta)$ defined by:

$$F(\beta) = \frac{|\{(i,j)|\tilde{\hat{\mathbf{\Omega}}}(i,j) \geqslant \beta, i\neq j, i=1,\ldots,n, j=1,\ldots,n\}|}{|\{(i,j)|i\neq j, i=1,\ldots,n, j=1,\ldots,n\}|},$$
$$0 \leqslant \beta \leqslant 1, \quad \tilde{\hat{\mathbf{\Omega}}}(i,j) = \frac{|\hat{\mathbf{\Omega}}(i,j)|}{\sqrt{\hat{\mathbf{\Omega}}(i,i)}\sqrt{\hat{\mathbf{\Omega}}(j,j)}},$$

where n is the dimensionality of the entire principal component space, $|\{\cdot\}|$ is the size of $\{\cdot\}$ and $\tilde{\hat{\mathbf{\Omega}}}(i,j)$ is the absolute standard correlation value between $\xi_{\mathbf{x}}^i$ and $\xi_{\mathbf{x}}^j$.

The curve of the value of $F(\beta)$ as a function of β has been shown in Figs. 12 and 13 on FERET and CMU PIE, respectively, where three training samples are used for each class on FERET and six training samples are used for each class on CMU PIE. We observe that on FERET, $F(\beta)=0.2925\%$ when $\beta=0.09959$ and $F(\beta)=0.006176\%$ when $\beta=0.2015$; on CMU, $F(\beta)=0.3002\%$ when $\beta=0.102$ and $F(\beta)=0.008472\%$ when $\beta=0.2513$. This shows that it would be quite a low probability for the absolute standard correlation value $\tilde{\hat{\mathbf{\Omega}}}(i,j)$, $i\neq j$ to get a high value. It means it has an extremely high probability that the correlation between $\xi_{\mathbf{x}}^i$ and $\xi_{\mathbf{x}}^j$ is very low when $i\neq j$.

In conclusion, the experiment shows that $\xi_{\mathbf{x}}^1,\ldots,\xi_{\mathbf{x}}^n$ are almost uncorrelated each other because of the extremely low correlation values between them. As we always do not have sufficient samples to tackle the ill-posed estimation problem when dealing with high-dimensional data, it is a practical and also reasonable way to hold this assumption for performing regularized estimation and model the perturbation covariance matrix using Eq. (21) and its further reduced form Eq. (22).

References

[1] R.A. Fisher, The statistical utilization of multiple measurements, Ann. Eugen. 8 (1938) 376–386.

[2] D.L. Swets, J. Weng, Using discriminant eigenfeatures for image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836.

[3] A.R. Webb, Statistical Pattern Recognition, second ed., Wiley, UK, 2002.

[4] W. Zhao, R. Chellappa, J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. (2003) 399–458.

[5] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[6] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, A new LDA-based face recognition system, which can solve the small sample size problem, Pattern Recognition 33 (10) (2000) 1713–1726.

[7] H. Cevikalp, M. Neamtu, M. Wilkes, A. Barkana, Discriminative common vectors for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005) 4–13.

[8] R. Huang, Q. Liu, H. Lu, S. Ma, Solving the small sample size problem in LDA, ICPR 2002.

[9] D.Q. Dai, P.C. Yuen, Regularized discriminant analysis and its application to face recognition, Pattern Recognition 36 (2003) 845–847.

[10] Z.-Q. Hong, J.-Y. Yang, Optimal discriminant plane for a small number of samples and design method of classifier on the plane, Pattern Recognition 24 (1991) 317–324.

[11] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition, Pattern Recognition Lett. 26 (2) (2005) 181–191.

[12] P. Zhang, J. Peng, N. Riedel, Discriminant analysis: a least squares approximation view, CVPR 2005.

[13] W. Zhao, R. Chellappa, P.J. Phillips, Subspace linear discriminant analysis for face recognition, Technical Report CAR-TR-914, CS-TR-4009, University of Maryland at College Park, USA, 1999.

[14] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data—with application to face recognition, Pattern Recognition 34 (2001) 2067–2070.

[15] J.P. Ye, Q. Li, LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation, Pattern Recognition 37 (4) (2004) 851–854.

[16] J. Duchene, S. Leclercq, An optimal transformation for discriminant and principal component analysis, IEEE Trans. Pattern Anal. Mach. Intell. 10 (6) (1988) 978–983.

[17] Z. Jin, J.Y. Yang, Z.S. Hu, Z. Lou, Face recognition based on the uncorrelated discriminant transformation, Pattern Recognition 34 (2001) 1405–1416.

[18] H. Xiong, M.N.S. Swamy, M.O. Ahmad, Two-dimensional FLD for face recognition, Pattern Recognition 38 (2005) 1121–1124.

[19] J. Yang, D. Zhang, X. Yong, J.-y. Yang, Two-dimensional discriminant transform for face recognition, Pattern Recognition 38 (2005) 1125–1129.

[20] J. Ye, R. Janardan, Q. Li, Two-dimensional linear discriminant analysis, NIPS 2004.

[21] W.-S. Zheng, J.H. Lai, S.Z. Li, 1D-LDA versus 2D-LDA: when is vector-based linear discriminant analysis better than matrix-based?, Pattern Recognition 41 (7) (2008) 2156–2172.

[22] J.R. Price, T.F. Gee, Face recognition using direct, weighted linear discriminant analysis and modular subspaces, Pattern Recognition 38 (2005) 209–219.

[23] J.H. Friedman, Regularized discriminant analysis, J. Am. Stat. Assoc. 84 (405) (1989).

[24] J.P. Hoffbeck, D.A. Landgrebe, Covariance matrix estimation and classification with limited training data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 763–767.

[25] H. Bensmail, G. Celeux, Regularized Gaussian discriminant analysis through eigenvalue decomposition, J. Am. Stat. Assoc. 91 (1996) 1743–1748.

[26] M. Zhu, A.M. Martinez, Subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (2006) 1274–1286.

[27] S.P. Lin, M.D. Perlman, A Monte Carlo comparison of four estimators of a covariance matrix, in: P.R. Krishnaiah (Ed.), Multivariate Analysis—VI: Proceedings of the Sixth International Symposium on Multivariate Analysis, Elsevier, Amsterdam, the Netherlands, 1985, pp. 411–429.

[28] S. Tadjudin, D.A. Landgrebe, Covariance estimation with limited training samples, IEEE Trans. Geosci. Remote Sensing 37 (4) (1999) 2113–2118.

[29] C.E. Thomaz, D.F. Gillies, R.Q. Feitosa, A new covariance estimate for Bayesian classifiers in biometric recognition, IEEE Trans. Circuits Syst. Video Technol. 14 (2) (2004) 214–223.

[30] P.W. Wahl, R.A. Kronmall, Discriminant functions when covariances are equal and sample sizes are moderate, Biometrics 33 (1977) 479–484.

[31] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 103–108.

[32] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 228–233.

[33] J. Yang, J.Y. Yang, Why can LDA be performed in PCA transformed space? Pattern Recognition 36 (2) (2003) 563–566.

[34] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1103.

[35] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1619.

[36] J. Yang, D. Zhang, J.-y. Yang, Median LDA: A robust feature extraction method for face recognition, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC2006), Taiwan.

[37] R. Herbrich, Learning Kernel Classifiers Theory and Algorithms, MIT Press, Cambridge, Massachusetts, London, England, 2002.

[38] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.

[39] M. Loog, R.P.W. Duin, Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 732–739.

About the Author—WEI-SHI ZHENG was born in Canton (Guangzhou), China, in 1981. He has recently received his Ph.D. degree in Applied Mathematics at Sun Yat-Sen University in China. He joined Queen Mary, University of London as a postdoctoral research assistant in August 2008. He is now working on the European SAMURAI Research Project with Prof. Gong Shaogang and Dr. Xiang Tao. Prior to that, he received his B.S. degree in both mathematics and computer science at Sun Yat-sen University in 2003. From April 2006 to October 2006, he was a visiting student working with Prof. Li Stan Z. at the Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. In 2007, he was an exchanged research student working with Prof. Yuen Pong C. at Hong Kong Baptist University from May 16 to November 15. Dr. Zheng is a member of the IEEE. His current research interests are in object categorization and semi-supervised learning; he is also interested in techniques for discriminant/sparse feature extraction and dimension reduction, kernel methods in machine learning, and face image analysis..

About the Author—J.H. LAI was born in 1964. He received the M.Sc. degree in applied mathematics in 1989 and the Ph.D. degree in mathematics in 1999 from Sun Yat-sen University, Guangzhou, China. He joined Sun Yat-sen University in 1989, where currently, he is a Professor with the Department of Electronics and Communication Engineering, School of Information Science and Technology. He has published over 50 papers in the international journals, book chapters, and conferences. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelets and their applications. Dr. Lai had successfully organized the International Conference on Advances in Biometric Personal Authentication' 2004, which was also the Fifth Chinese Conference on Biometric Recognition (Sinobiometrics'04), Guangzhou, in December 2004. He has taken charge of more than four research projects, including NSFC (number 60144001, 60 373 082, 60675016), the Key (Key grant) Project of Chinese Ministry of Education (number 105 134), and NSF of Guangdong, China (number 021 766, 06023194). Dr. Lai has published over 60 papers and he serves as a board member of the Image and Graphics Association of China and also serves as a board member and secretary-general of the Image and Graphics Association of Guangdong.

About the Author—PONG C YUEN received his B.Sc. degree in Electronic Engineering with first class honours in 1989 from City Polytechnic of Hong Kong, and his Ph.D. degree in Electrical and Electronic Engineering in 1993 from The University of Hong Kong. He joined the Department of Computer Science, Hong Kong Baptist University in 1993 as an Assistant Professor and currently is a Professor.

Dr. Yuen was a recipient of the University Fellowship to visit The University of Sydney in 1996. He was associated with the Laboratory of Imaging Science and Engineering, Department of Electrical Engineering and worked with Prof. Hong Yan. In 1998, Dr. Yuen spent a six-month sabbatical leave in the University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland at college park. He was associated with the Computer Vision Laboratory, CFAR and worked with Prof. Larry Davis. From June 2005 to January 2006, he was a visiting professor in GRAVIR laboratory (GRAphics, VIsion and Robotics) of INRIA Rhone Alpes, France. He was associated with PRIMA Group and work with Prof. James Crowley. Dr. Yuen was the director of Croucher Advanced Study Institute (ASI) on biometric authentication in 2004 and was the director of Croucher ASI on Biometric Security and Privacy in 2007. Dr. Yuen has been actively involved in many international conferences as an organizing committee and/or technical program committee member. Recently, he was the track co-chair of International Conference on Pattern Recognition 2006. Dr. Yuen is an editorial board member of Pattern Recognition.

Dr. Yuen's current research interests include human face processing and recognition, biometric security and privacy, context modeling and learning for human activity recognition.

About the Author—STAN Z. LI received his Ph.D. degree from Surrey University, UK. He is currently a professor at the National Laboratory of Pattern Recognition (NLPR), the director of Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA); and co-director of Joint Laboratory for Intelligent Surveillance and Identification in Civil Aviation (CASIA-CAUC). He worked at Microsoft Research Asia as a researcher from 2000 to 2004. Prior to that, he was an associate professor at Nanyang Technological University, Singapore. His research interest includes pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance. He has published over 200 papers in international journals and conferences, and authored and edited 5 books including "Markov Random Field Modeling in Image Analysis" (Springer, 1st edition 1995, 2nd edition 2001, 3rd edition 2008). He is currently an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and has been actively participating in organizing a number of international conferences and workshops in the fields of his research interest. Stan Z. Li is an expert in face recognition, biometrics and intelligent video surveillance. The Eye-CU face recognition system he developed at Microsoft Research Asia was demonstrated by Bill Gate on a CNN interview. He has been leading several national and international collaboration projects in biometrics and intelligent video surveillance. The AuthenMetric face recognition system and intelligent video surveillance system have been deployed in many applications. He acted as the program chair for the Asian Biometrics Forum 2006 and a co-chair for the International Conference on Biometrics 2007 and 2009. He delivered a speech on Biometrics in China, on behalf of the China National Body, at the 2006 ISO/IEC JTC1 SC37 meeting in London. He co-edited Handbook of Face Recognition (Springer, 2005), and is acting as the editor-in-chief for the Encyclopedia of Biometrics (Springer Reference Work, to be published in 2009).