Sampling Representative Examples for Dimensionality Reduction and Recognition – Bootstrap Bumping LDA

Hui Gao and James W. Davis

Dept. of Computer Science and Engineering
The Ohio State University, 2015 Neil Ave,
Columbus, OH 43220, USA
{gaoh, jwdavis}@cse.ohio-state.edu

Abstract. We present a novel method for dimensionality reduction and recognition based on Linear Discriminant Analysis (LDA), which specifically deals with the Small Sample Size (SSS) problem in Computer Vision applications. Unlike the traditional methods, which impose specific assumptions to address the SSS problem, our approach introduces a variant of bootstrap bumping technique, which is a general framework in statistics for model search and inference. An intermediate linear representation is first hypothesized from each bootstrap sample. Then LDA is performed in the reduced subspace. Lastly, the final model is selected among all hypotheses for the best classification. Experiments on synthetic and real datasets demonstrate the advantages of our Bootstrap Bumping LDA (BB-LDA) approach over the traditional LDA based methods.

1 Introduction

As a statistical method for dimensionality reduction and classification [1], Linear Discriminant Analysis (LDA) has been widely employed in Computer Vision research (e.g., face and gait recognition [2, 3, 4, 5]). Since LDA assumes multiple Gaussians with equal covariance, its success largely depends on accurate estimates of the model parameters (class means and common covariance). However in most Computer Vision applications, the sample size N is relatively small in comparison to the input dimension D. The traditional Maximum Likelihood (ML) estimates show poor convergence to the true parameters due to the curse of the dimensionality. Furthermore, when $N < D$, the ML estimate of the common covariance $\hat{\Sigma}$ is even singular (the LDA solution is underconstrained due to the non-existence of $\hat{\Sigma}^{-1}$). These two issues together constitute the so-called Small Sample Size (SSS) problem in LDA.

The traditional LDA methods [6, 2, 3, 7, 8] focus only on the second issue of a singular $\hat{\Sigma}$, but ignore the accurate estimate of the true model parameters. Even if $N > D$, as long as N is not much larger than D, the SSS problem persists. From this point of view, the dual impact of the SSS problem is crucial to the success of LDA in Computer Vision applications.

In this work, we propose to deal with the SSS problem from a more general aspect with the goal of accurately estimating the model parameters. Instead of imposing explicit assumptions to simply invert the singular $\hat{\Sigma}$, we introduce a variant of a general

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part III, LNCS 3953, pp. 275–287, 2006.

statistical framework, bootstrap bumping, which creates a hypothesis from each bootstrap sample (a subset of examples) and selects the best model according to a target criteria. The original bumping technique was developed in [9] for finding better local minima, resistant fitting, and optimization under constraints. We develop the idea to deal with the SSS problem by hypothesizing an intermediate linear representation from each bootstrap sample and choosing the final model (representation) with the best recognition performance. This extension not only has the same asymptotical property as the original bumping procedure, but now improves the estimation accuracy and implicitly handles the singularity problem of $\hat{\Sigma}$ in the SSS problem. We present experiments on synthetic and real datasets to clearly show the advantages of our approach over the traditional LDA methods.

In the remainder of this paper, we first discuss the background and related work of LDA and bootstrap bumping in Sect. 2. Then we describe our proposed approach in Sect. 3, which specifically deals with the SSS problem in LDA. Lastly, experimental results are presented in Sect. 4, followed by conclusions in Sect. 5.

2 Background and Related Work

There are two different perspectives looking at LDA. Fisher's LDA is defined by maximizing the ratio of the between-class and within-class scatter matrices (S_b and S_w) in a linear feature space [10, 11]. In Bayesian decision theory, LDA is defined for the case of multiple Gaussians with equal covariance. The two approaches were shown to be equivalent in [12] with S_w being the ML estimate $\hat{\Sigma}$ and S_b being derived from the ML estimates of the class means. The mathematical description of both approaches can be found in detail in [13], which we omit here due to space constraints.

2.1 LDA and the SSS Problem

Although well-grounded in theory, LDA faces the challenge of the SSS problem in real applications. Traditional methods only aim to solve the singularity problem of $\hat{\Sigma}$ by imposing specific assumptions to simply invert $\hat{\Sigma}$.

The simple approach PINV-LDA [6] substitutes the inverse operation with pseudoinverse. The two-stage method PCA+LDA [2] projects the data in the nearly complete PCA subspace to make the $\hat{\Sigma}$ projection just full rank. However, with a small number of examples, $\hat{\Sigma}$ is unstable especially in those components with small eigenvalues which are mostly emphasized in the inverse operation. Both methods of PINV-LDA and PCA+LDA are sensitive to noise and small perturbations.

As one improvement, Enhanced Fisher's Linear Discriminant (EFLD) [3] varies the number of PCA components to regulate the projection of $\hat{\Sigma}$. This assumes that the small components are not informative for classification, which may impose a performance limitation. Another approach Direct LDA (D-LDA) [7] assumes the null space of S_b contains no useful information for recognition. However, as shown in our prior work [14], D-LDA is equivalent to directly taking the linear space of class means as the LDA solution. It has severe limitations by ignoring the common covariance estimate $\hat{\Sigma}$ (or S_w). Lastly, $\hat{\Sigma}$ can be modified to avoid the singularity problem, such as $\hat{\Sigma} + \sigma I$

in Regularized LDA (R-LDA) [8][1]. With σ usually being a small scalar, R-LDA heavily relies on the small components and even null components for recognition, which is neither stable nor supported by the existing examples. For $\sigma = inf$, R-LDA is equivalent to D-LDA by ignoring $\hat{\Sigma}$. Furthermore, R-LDA is computationally inefficient for a large input dimension D since the full-rank matrix $\hat{\Sigma} + \sigma I$ is of size $D \times D$ and is to be inverted in LDA.

Traditional LDA methods only focus on the singularity problem of $\hat{\Sigma}$. Systematic attempts to reduce the variance of the ML estimates (for both class means and common covariance) in the general SSS problem have not yet been reported. We address this issue in our proposed framework of Bootstrap Bumping LDA.

Additionally there are approaches to address the model limitations of LDA, such as methods to extract non-linear features in Quadratic Discriminant Analysis (QDA) [13] for multiple Gaussians with non-equal covariance, kernel-based Generalized LDA (GLDA) [15], and Locally Linear Discriminant Analysis (LLDA) [16]. Since more examples are usually required to constrain more complex solutions, these methods are even more sensitive to the SSS problem. As a hybrid model of LDA and QDA, Oriented Discriminant Analysis (ODA) [17] assumes the same as QDA of multiple Gaussians with non-equal covariance, but extracts linear features by maximizing the Kullback-Liebler divergence between classes. However its explicit explanation remains unclear in Bayesian decision theory since quadratic features are inherently required under the model assumption. As another modification, Optimal Linear Representation [18] allows classifiers (e.g., k-Nearest Neighbor) other than thresholding (assumed by LDA) by searching the solution space (a set of linear subspaces, or Grassmann manifold) with regard to a searching strategy. But this heuristic approach lacks theoretical support from Bayesian decision theory. It is computationally expensive, as Markov Chain Monte Carlo (MCMC) simulation is often employed, and it is even doubtful whether such a search is bounded or stable in a high dimensional space with few examples.

2.2 Bootstrap Methods

The general bumping procedure was proposed in [9] as a method for model search and inference. It is based on bootstrap resampling theory [19], which was originally used for assessing the statistical accuracy of an estimator. A "bootstrap sample" is a "subset of examples" randomly drawn *with replacement* from the original set of training examples. It was shown that the empirical distribution of bootstrap samples can be used to approximate the sampling distribution of random variables (e.g., variance of an estimator) to be estimated from the observed data. Additionally, recent research demonstrated that the bootstrap technique can be employed to improve the accuracy of an estimator, such as *bagging* [20], *boosting* [21, 22] (with an enhanced version called AdaBoosting [23, 24] which employs adaptive sampling and weighted voting), and *bumping* [9]. By averaging the estimates from multiple bootstrap samples, *bagging* produces a new estimator, which often has a smaller variance. In comparison, the *boosting* method improves the classification performance by combining multiple weak learners, individually trained

[1] The original idea was to smoothly blend LDA with Quadratic Discriminant Analysis (QDA) by adding the common covariance (scaled by σ) to the individual covariance of each class. Although not explicitly described in [8], R-LDA is often referred to $\hat{\Sigma} + \sigma I$ in the literature.

from a subset of examples. However, if we desire a single LDA classifier or a set of LDA linear features for dimensionality reduction, the bagged (averaged) linear classifier from subsets may not perform well, and the boosted classifier results in complex decision boundaries, which is non-linear and is not applicable for dimensionality reduction. In this sense, both bagged and boosted LDA [25, 26] are no longer true "LDA".

However, in our proposed approach, the *bumping* procedure [9] follows the paradigm of hypothesis and test. Bootstrap samples are used to provide candidate models. The procedure then selects the model which best explains the observed data according to a target criteria. The method reduces the variance of the original estimates, while preserving the same structure and interpretation. This ideally suites our need to address the SSS problem in LDA.

3 Bootstrap Bumping LDA (BB-LDA)

The original bumping procedure [9] directly hypothesizes a model from each bootstrap sample and selects the best model for a target criteria. However, this approach is not directly applicable to the SSS problem in LDA. Because each bootstrap sample contains even fewer examples, the SSS problem is more problematic for the LDA model directly trained/estimated from bootstrap samples. Furthermore, the singularity problem of $\hat{\Sigma}$ in LDA is not yet addressed in the original bumping procedure.

Instead we propose a *new* bumping procedure called Bootstrap Bumping LDA (BB-LDA). The approach first hypothesizes an intermediate linear representation from each bootstrap sample. Then all of the training examples are projected into the representation space and analyzed by the classic LDA. The new procedure not only has the same asymptotic property of convergence as original bumping, but now avoids the singularity problem of $\hat{\Sigma}$ and improves the estimation accuracy of model parameters in the SSS problem. Our approach is significant in that it addresses the dual aspects of the SSS problem in a general statistical framework without imposing specific assumptions (as the traditional methods). It also preserves LDA interpretation by avoiding averaging (bagging) or voting (boosting). We begin with a description of the general bumping procedure in Sect. 3.1 and present our extension in Sect. 3.2.

3.1 Bootstrap Bumping

Let $\mathbf{z} = (z_1, z_2, \cdots, z_N)$ be the set of all labeled training examples. Assume a data model depends on a set of parameters θ, which is to be estimated by minimizing a *target* criteria R as

$$\hat{\theta} = \text{argmin}_\theta \; R(\mathbf{z}, \theta). \tag{1}$$

The criteria R can be of any general form, such as median squared error for linear regression, or the Maximum Likelihood (ML) estimates of the model parameters, which have closed-form solutions. Ultimately, minimizing R obtains the target estimation $\hat{\theta}$ from the input data $\mathbf{z}$

Suppose there is another *working* criteria R_0, which may be more convenient for minimization (e.g., replacing least median square with least mean square). At a particular sampling rate/ratio α, each bootstrap sample $\mathbf{z}^{*1}, \mathbf{z}^{*2}, \cdots, \mathbf{z}^{*B}$ is randomly drawn

from $\mathbf{z}$ with replacement (each sample has αN training examples). The estimate of θ via R_0 from each bootstrap sample is

$$\hat{\theta}^{*b} = \text{argmin}_\theta \; R_0(\mathbf{z}^{*b}, \theta). \tag{2}$$

The original bumping procedure [9] chooses $\hat{\theta}^{\mathcal{BB}}$ as the value among the $\hat{\theta}^{*b}$ which has the smallest value in the target criteria $R(\mathbf{z}, \theta)$ for the entire dataset $\mathbf{z}$:

$$\hat{\theta}^{\mathcal{BB}} = \hat{\theta}^{*\hat{b}}, \text{ where } \hat{b} = \text{argmin}_b \; R(\mathbf{z}, \hat{\theta}^{*b}) \tag{3}$$

The working criteria R_0 may be the same as the target criteria R, in which case the bumping procedure simply estimates suboptimal parameters $\hat{\theta}^{*b}$ from each bootstrap sample (a subset of training examples) and selects the best $\hat{\theta}^{*b}$ over all hypothesized candidates. This has been shown in [9] to be useful for finding a better local minima. For different working and target criteria [9], the bumping procedure can also be used for robust fitting (with R_0 as the outlier-free version of R) and constrained optimization (R_0 as the unconstrained version of R).

Furthermore, the working criteria R_0 needs to be "compatible" with the target criteria R in order for the bumping estimate $\hat{\theta}^{\mathcal{BB}}$ to asymptotically converge to the true model parameters θ. For the same criteria R_0 and R, it has been proven in [9] that the bumping procedure preserves the property of asymptotic convergence. For different criteria R_0 and R, compatibility should be carefully examined by considering the asymptotic behavior of the procedure. Otherwise, the bumping procedure only provides an approximation of R with a simple form R_0 largely for the ease of computation.

3.2 Proposed Approach – Bootstrap Bumping LDA

The original bumping procedure was not designed to handle the SSS problem. With regards to LDA, we choose the target criteria R as the ML solution, which measures the misclassification rate on $\mathbf{z}$ for linear decision boundaries θ. The minimization of R has a closed-form solution by first obtaining the ML estimates of LDA from $\mathbf{z}$ and then calculating the corresponding decision boundaries $\hat{\theta}$ (the linear projections and thresholds). If we employ the same working criteria $R_0 = R$, the original bumping procedure hypothesizes linear decision boundaries $\hat{\theta}^{*b}$ from each bootstrap sample $\mathbf{z}^{*b}$. However, since each bootstrap sample $\mathbf{z}^{*b}$ has fewer examples than $\mathbf{z}$, when there are not enough examples, the estimate $\hat{\theta}^{*b}$ is even more unstable than the original $\hat{\theta}$. The impact of the SSS problem is magnified, not suppressed.

To deal with this issue, instead of directly estimating θ, we propose to first hypothesize an intermediate representation space $\hat{L}^{*b}$ from each bootstrap sample $\mathbf{z}^{*b}$ as

$$\hat{L}^{*b} = \text{argmin}_L \; R_{rep}(\mathbf{z}^{*b}, L). \tag{4}$$

Here the new *working* criteria R_{rep} measures the capacity of a given representation L (e.g., linear, quadratic, etc.) for the bootstrap sample $\mathbf{z}^{*b}$, which we call the *representation* criteria. We want to choose the representation with minimum capacity (the simplest representation), which still faithfully reconstructs the bootstrap sample and is compatible with the model assumption. With regards to LDA, a linear subspace defined by $\mathbf{z}^{*b}$

is minimum in terms of capacity among all compatible representations. Therefore we can directly replace Eqn. 4 with

$$\hat{L}^{*b} = LinearSpace(\mathbf{z}^{*b}). \tag{5}$$

For other models, the representation should be chosen accordingly. For example, for QDA a quadratic representation should be hypothesized from each bootstrap sample.

Then we follow the similar bumping procedure. we evaluate the discrimination performance of the hypothesized representation $\hat{L}^{*b}$ over the entire dataset $\mathbf{z}$ and choose the representation with the minimum misclassification rate as in

$$\hat{L}^{\mathcal{BB}-\mathcal{LDA}} = \hat{L}^{*\hat{b}}, \text{ where } \hat{b} = \text{argmin}_b \, R_{dis}(\mathbf{z}, \hat{L}^{*b}) \tag{6}$$

The new *target* criteria R_{dis} measures the misclassification rate of $\mathbf{z}$ with regard to the representation space $\hat{L}^{*b}$, which we call the *discrimination* criteria. The target criteria R_{dis} can be easily evaluated using the best estimated model parameters $\hat{\theta}^{*b}$ based on the representation $\hat{L}^{*b}$

$$R_{dis}(\mathbf{z}, \hat{L}^{*b}) = R(\mathbf{z}, \hat{\theta}^{*b}), \text{ where} \tag{7}$$

$$\hat{\theta}^{*b} = \text{argmin}_\theta \, \tilde{R}(\mathbf{z}, \theta; \hat{L}^{*b}). \tag{8}$$

As a constrained version of the original bumping criteria R in the representation space $\hat{L}^{*b}$, the modified criteria $\tilde{R}$ is equivalent to first projecting $\mathbf{z}$ into $\hat{L}^{*b}$ (e.g., correlating with a linear basis in LDA), estimating the model parameters (e.g., ML), and lastly reconstructing the parameters back to the original D-dimensional space (e.g., multiplying the feature vectors with the basis).

Lastly, we obtain the LDA solution of BB-LDA as the corresponding model estimates for the selected representation space

$$\hat{\theta}^{\mathcal{BB}-\mathcal{LDA}} = \hat{\theta}^{*\hat{b}}. \tag{9}$$

In essence, the approach seeks out the key prototype examples that best represent the space of $\mathbf{z}$ for the purpose of discrimination. The BB-LDA algorithm is summarized in Alg. 1. For any new example z_{new}, it can then be classified by projecting it onto the reconstructed feature space and thresholding.

Our proposed approach addresses the SSS problem in a general statistical framework. At a particular sampling ratio α, only a portion of examples are used to hypothesize a representation, which can ensure $\hat{\Sigma}$ being full rank in the projection space $\hat{L}^{*b}$ for the entire dataset $\mathbf{z}$. Since duplicate examples do not affect the representation, bootstrap samples are drawn at a fixed size αN from $\mathbf{z}$ *without replacement* in BB-LDA for the ease of analysis and implementation. The prior probability of each class is also maintained in sampling to ensure a fair representation. Furthermore because LDA is invariant to the basis selection, a non-orthonormal basis T_b is used for simplicity, which is equivalent to linearly correlating the entire dataset with examples in each bootstrap sample. The smaller the sampling ratio α, the more compact the representation, and the more examples left to generalize the model for discrimination. However, too few examples negatively affect the representation power, which may in turn limit the upper bound for

Algorithm 1. BB-LDA Algorithm

1: Randomly draw B bootstrap samples $\mathbf{z}^{*1}, \mathbf{z}^{*2}, \cdots, \mathbf{z}^{*B}$ from $\mathbf{z}$ at the sampling ratio α.
2: **for** $b = 1$ to B **do**
3: Let $A_b = [z_1^{*b}, z_2^{*b}, \cdots, z_k^{*b}]$ be one basis of the linear subspace $\hat{L}^{*b}$.
4: Project $\mathbf{z}$ in A_b as $y_b = A_b^T \mathbf{z}$. Run LDA with ML estimates on y_b to obtain the model parameters, including the feature vector(s) w_b and the threshold(s) t_b.
5: Calculate the misclassification rate on y_b based on the estimated model parameters.
6: **end for**
7: Choose the representation A_b which has the minimum misclassification rate. Obtain the BB-LDA solution $\hat{\theta}^{\mathcal{BB}-\mathcal{LDA}}$ by reconstructing the feature vectors $A_b w_b$ and keeping the same threshold t_b.

discrimination. The application-dependent sampling ratio α can be determined through cross-validation to properly balance the representation and discrimination.

With regards to the the number of bootstrap samples B, the percentage of training examples p covered by all bootstrap samples is $p = 1 - (1 - \alpha)^B$. At a given sampling ratio α, B can be calculated for a specific coverage (e.g., $p = 99.9\%$) with

$$B = log(1 - p)/log(1 - \alpha). \tag{10}$$

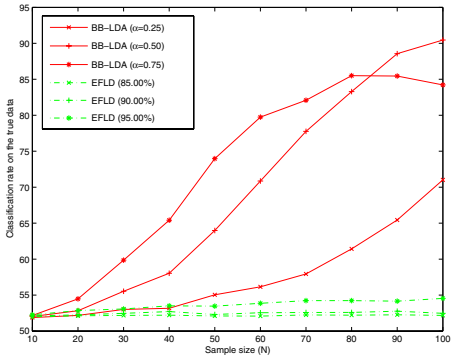
While the traditional subspace LDA approaches (e.g., PINV-LDA, PCA+LDA) have the time complexity of $O(N^2 D)$ in the SSS problem, BB-LDA has the time complexity of $O(B\alpha N^2 D)$. From Eqn. 10, the worse case time complexity of BB-LDA occurs at $O(-log(1 - p)N^2 D)$ when $\alpha \to 0$, which is on the same order as the traditional subspace LDA [6, 2, 3]. Additionally, it is possible to reduce the computational cost of BB-LDA with a smaller coverage p, which may be useful for extremely large datasets.

Since different *working* (R_{rep}) and *target* (R_{dis}) criteria are used for representation and discrimination, according to the bumping theory [9], R_{rep} needs to be "compatible" with R_{dis} in order for our new procedure to asymptotically converge to the true parameters. This can be proved by considering the compatibility between a linear representation and LDA. At a fixed sampling ratio α, when the number of representative examples $\alpha N > D$ (as in BB-LDA), the representation space is even larger than the original input space (assuming linear independence among examples in each bootstrap sample). Estimating an LDA model in each hypothesized representation in Eqn. 8 is equivalent to directly applying ML in the original input space. Thus the bumping procedure is equivalent to LDA with ML estimates when $N > D/\alpha$. Because of the asymptotic convergence property of the ML estimates, this proves the compatibility of our *working* criteria R_{rep} and *target* criteria R_{dis}.

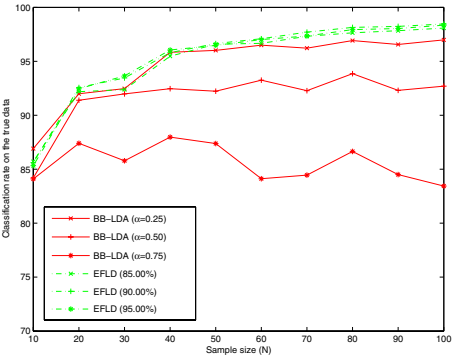
Lastly, the bootstrap sampling process is not limited to be uniform. Parameterized bootstrapping [27] can be utilized to accommodate the underlying structure of the data. Other extensions, such as employing clustering or domain knowledge for bootstrap resampling, are possible.

4 Experiments

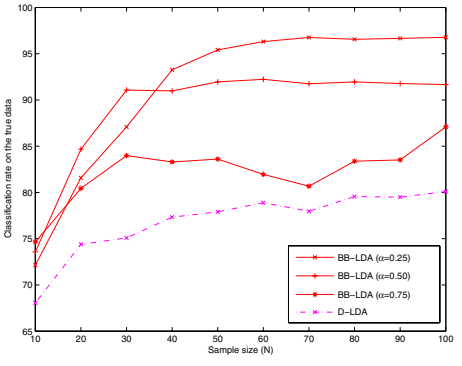
We evaluated the performance of BB-LDA with both synthetic and real datasets in comparison to traditional LDA methods in dealing with the SSS problem.



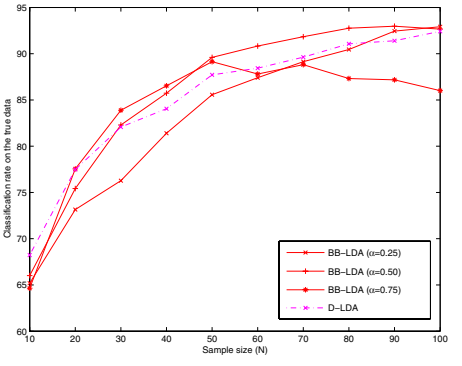
(a) Bad case for EFLD



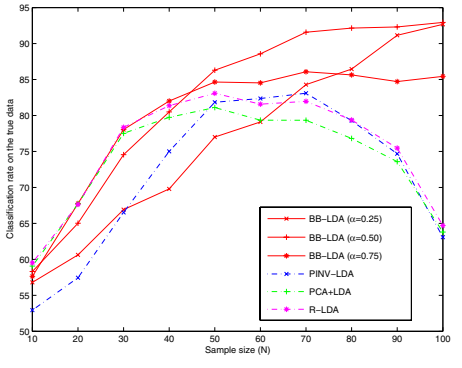
(b) Good case for EFLD



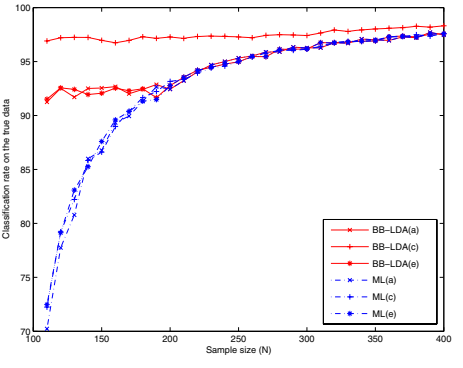
(c) Bad case for D-LDA



(d) Good case for D-LDA



(e) Sensitivity to noise



(f) Larger sample sizes

Fig. 1. Results of synthetic experiments. With the best $\alpha \in [0.25, 0.5, 0.75]$, BB-LDA outperformed traditional LDA methods when their assumptions were intentionally violated in (a), (c), and (e), and yielded comparable performance when the assumptions were satisfied in (b) and (d). As shown in (f), BB-LDA converges to the ML estimate with a large enough sample size.

4.1 Results on Synthetic Data

In our synthetic experiments, two Gaussians with equal covariance were simulated with equal priors in a $D = 100$ dimensional space for a range of sample sizes $N = [10 : 10 : 400]$. The difficulty of the classification was controlled using a fixed Fisher ratio of 4, which corresponds to a 97.7% Bayesian classification rate. Each configuration (class means and common covariance) was simulated 25 times to report the average recognition rate of the model with regards to the ground-truth data. We chose the percentage coverage $p = 99.9\%$ to determine the number of bootstrap samples, which achieves good utilization of training examples and reasonable computational efficiency.

We first looked at the case of $N \leq 100$ (singular $\hat{\Sigma}$), which was previously focused on by the traditional methods (PINV-LDA, PCA+LDA, EFLD, R-LDA, and D-LDA). As shown in Fig. 1a, the EFLD recognition rates were hardly better than 50% for 3 selected percentage fits (85%, 90%, and 95%) of the simulated data, when the true feature vector lies outside the major PCA components (90% fit of the true data variance). This is because EFLD assumes no information in the small components and discards them to constrain the LDA solution. Similarly, D-LDA showed low performance in Fig. 1c when a large portion of the true feature vector resides in the null space of S_b (class means). For the remaining methods, PINV-LDA, PCA+LDA, and R-LDA also performed poorly for a large N as they are sensitive to noise and small perturbations due to the over-emphasis of small components in their solutions (Fig. 1e).

As a comparison, at an appropriate sampling ratio α, BB-LDA outperformed the traditional methods in all the above cases (see Fig. 1a,c,e). Furthermore, when the model assumptions of EFLD and D-LDA *were* satisfied as shown in Fig. 1b and 1d, BB-LDA still yielded comparable performance to the two methods. The valid case of PINV-LDA, PCA+LDA, and R-LDA is not available due to their unstable nature.

Then we studied the performance of BB-LDA in handling the SSS estimation problem for the case (a), (c), and (d) with $100 < N \leq 400$, which has enough examples to avoid a singular $\hat{\Sigma}$. The sampling ratio was selected for the best average recognition rate in the previous range of $N \leq 100$ with α at 0.5, 0.25, and 0.5. As shown in Fig. 1f, BB-LDA outperformed classic LDA in the lower end of the range of N, due to relatively few examples for the ML estimates to converge. In the higher end with enough examples, BB-LDA showed the trend of convergence to ML. The results demonstrate BB-LDA as a general method to deal with the SSS problem in various cases.

4.2 Results on Real Data

In our real experiments, we explored 3 datasets frequently used in Computer Vision research for face and gait recognition: Yale face database [2], ORL face dataset [28], and the CMU gait database [29]. For each dataset, images were first aligned to control position and scaling. Then they were down-sampled and tightly cropped to the region of interest as shown in Fig. 2. For the gait database, two different types of MHI (overlay of silhouette images with timestamps represented in pixel intensity) [30] templates were created, which correspond the stride opening and closing phase of a walking cycle (Fig. 2c and 2d). All traditional methods used in Sect. 4.1 were evaluated except R-LDA due to its inherit high computational complexity for a large input dimension (e.g., $D = 1600$ for images in the Yale face database). Cross-validation was employed to determine the

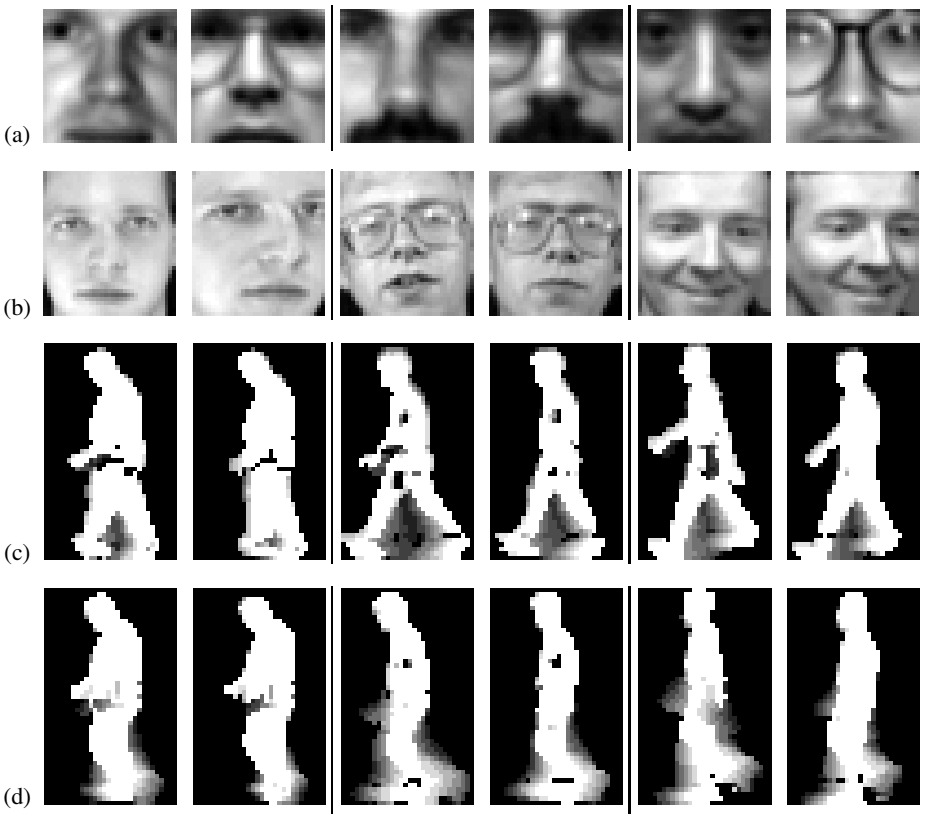


Fig. 2. Sample images of 3 datasets. (a) Yale face database (15 subjects, glasses vs. no glasses). (b) ORL face dataset (40 subjects). (c) CMU Gait database (25 subjects, fast vs. slow walk) in Type-1 MHI representation. (d) Corresponding Type-2 MHI Gait representation.

optimal model parameters of BB-LDA (the sampling ratio $\alpha \in [0.1 : 0.1 : 0.9]$ and the best representation/classifier) and EFLD (the number of PCA components) with each time 10% of examples drawn for testing. The same bootstrap coverage $p = 99.9\%$ was chosen as in Sect. 4.1.

The Yale face dataset includes 15 subjects and 11 images of each person across various conditions (e.g., lighting, expressions, etc.). In addition to face recognition, we examined the task of distinguishing people with glasses from people without glasses (36 with and 129 without), a much larger set than the case of 36 images studied in [2]. We then examined face recognition using the ORL face dataset with 40 subjects and 10 images per person. Lastly, we looked at the CMU gait database of 25 subjects with 16 cycles extracted for each person (8 slow and 8 fast). Both identity and walking speed recognition were performed over two types of MHI representation.

The comparative results of those experiments are summarized in Table 1. Since PINV-LDA and PCA+LDA mostly emphasize the small components, they are sensitive to noise and yielded lower recognition rates. By adjusting the number of PCA components, EFLD improved the performance of PCA+LDA and is the best among

Table 1. Classification results of different LDA-based algorithms. Our proposed BB-LDA approach outperformed the other traditional LDA methods.

	Yale - ID (11 sets)	Yale - Glasses (36 sets)	ORL - ID (10 sets)	CMU - ID (24 sets)		CMU - Speed (30 sets)	
				Type-1	Type-2	Type-1	Type-2
PINV-LDA	82.7	83.6	88.8	99.7	99.1	92.6	89.5
PCA+LDA	45.5	85.2	27.3	54.6	62.7	93.6	90.7
EFLD	90.6 (57 PCs)	89.7 (85 PCs)	92.3 (95 PCs)	100.0 (90 PCs)	99.7 (132 PCs)	97.0 (318 PCs)	95.3 (321 PCs)
D-LDA	70.3	72.0	79.8	77.8	76.3	77.4	65.8
BB-LDA	**93.9** ($\alpha = 0.3$)	**95.1** ($\alpha = 0.3$)	**95.5** ($\alpha = 0.2$)	**100.0** ($\alpha = 0.2$)	**100.0** ($\alpha = 0.2$)	**97.8** ($\alpha = 0.6$)	**97.1** ($\alpha = 0.4$)

all the traditional methods. But this assumes the small components contain no information for classification. Lastly, D-LDA imposes a significant performance limitation by constraining the feature vectors to be in the linear space of S_b (class means). As a comparison, our proposed BB-LDA approach gave the best classification rate in all test cases. Only in CMU-ID (Type-1), EFLD yielded the same classification of 100%, which is high due the simplicity of the task (MHI images of multiple cycles for one subject are highly similar).

The performance advantages of BB-LDA come from the employment of a general statistical framework of bootstrap bumping in dealing with the SSS problem. This avoids the explicit assumptions in the traditional methods. By sampling a subset of training examples to hypothesize a representation and selecting the best model for discrimination over the entire dataset, our approach is capable of improving the estimation accuracy in the SSS problem. The sampling ratio α provides a balance of examples for representation and discrimination. In our real experiments, a small α value was used in most cases, which suggests that only a few prototype examples were needed for representation, while the rest can be used for discrimination. Both synthetic and real experiments illustrated the advantages of BB-LDA in dealing with the SSS problem.

5 Conclusion

We presented a novel method of Bootstrap Bumping LDA (BB-LDA) to deal with the SSS problem in Computer Vision applications. The method hypothesizes candidate representations from each subset of examples (bootstrap sample) and tests over the entire dataset for the best classification. As a general statistical framework, our approach is capable of improving the estimation accuracy without imposing explicit assumptions. The method asymptotically converges to the true LDA solution given enough examples and outperforms the traditional LDA methods in dealing with the SSS problem. Both synthetic and real experiments on several popular datasets showed the advantages of our BB-LDA approach. In future work, we plan to address the model limitations of LDA with more complex representations (e.g., non-linear) and investigate other applications of BB-LDA (e.g., person detection).

Acknowledgments

This research was supported in part by the National Science Foundation under grant No. 0236653.

References

1. Jain, A., Duin, R., Mao, J.: Statistical pattern recognition: a review. IEEE Trans. Patt. Analy. and Mach. Intell. **22**(1) (2000) 4–37
2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Patt. Analy. and Mach. Intell. **19**(7) (1997) 711–720
3. Liu, C., Wechsler, H.: Enhanced Fisher linear discriminant models for face recognition. In: Proc. Int. Conf. Pat. Rec., IEEE (1998) 1368–1372
4. Cui, Y., Swets, D., Weng, J.: Learning-based hand sign recognition using SHOSLIF-M. In: Proc. Int. Conf. Comp. Vis., IEEE (1995) 631–636
5. Huang, P., Harris, C., Nixon, M.: Human gait recognition in canonical space using temporal templates. In: Proc. Vision Image Signal Process. Volume 146., IEE (1999) 93–100
6. Krzanowski, W., Jonathan, P., McCarthy, W., Thomas, M.: Discriminant analysis with singular covariance matrices:methods and applications to spectroscopic data. Applied Statistics **44** (1995) 101–115
7. Yu, H., Yang, J.: A direct LDA algorithm for high-dimensional data - with application to face recognition. Pattern Recognition **34** (2001) 2067–2070
8. Friedman, J.: Regularized discriminant analysis. J. Am. Statistical Assoc. **84**(405) (1989) 165–175
9. Tibshirani, R., Knight, K.: Model search by bootstrap "bumping". J. of Computational and Graphical Statistics **8**(4) (1999) 671–686
10. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7 Part II** (1936) 179–188
11. Rao, C.: The utilization of multiple measurements in problems of biological classification. J. Royal Statistical Soc., B **10** (1948) 159–203
12. Campbell, N.: Canonical variate analysis - a general model formulation. Australian J. Statistics **26** (1984) 86–96
13. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley & Sons, New York (2001)
14. Gao, H., Davis, J.: Why Direct LDA is not equivalent to LDA. to appear in Pattern Recognition (2006)
15. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation **12**(10) (2000) 2385–2404
16. Kim, T., Kittler, J.: Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. IEEE Trans. Patt. Analy. and Mach. Intell. **27**(3) (2005) 318–327
17. Torre, F., Kanade, T.: Oriented discriminant analysis (ODA). In: Brit. Mach. Vis. Conf. (2004) 132–141
18. Liu, X., Srivastava, A., Gallivan, K.: Optimal linear representations of images for object recognition. IEEE Trans. Patt. Analy. and Mach. Intell. **26**(5) (2004) 662–666
19. Efron, B.: Bootstrap methods: another look at the jackknife. Annals of Statistics **7** (1979) 1–26
20. Breiman, L.: Bagging predictors. Machine Learning Journal **24**(2) (1996) 123–140
21. Schapire, R.: The strength of weak learnability. Machine Learning **5**(2) (1990) 197–227

22. Freund, Y.: Boosting a weak learning algorithm by majority. Information and Computation **121**(2) (1995) 256–285
23. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Machine Learning: Proc. of the 13th Int. Conf. (1996) 148–156
24. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: Proc. Int. Conf. Comp. Vis. (2003) 734–741
25. Skurichina, M., Duin, R.: Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis & Applications **5** (2002) 121–135
26. Lu, X., Jain, A.K.: Resampling for face recognition. In: Int. Conf. on Audio and Video Based Biometric Person Auth. (2003) 869–877
27. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993)
28. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: 2nd IEEE Workshop on Applications of Computer Vision. (1994)
29. R.Gross, Shi, J.: The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (2001)
30. Davis, J., Bobick, A.: The representation and recognition of action using temporal templates. In: Proc. Comp. Vis. and Pattern Rec., IEEE (1997) 928–934