Population and Genetic Analysis

Semi-supervised feature extraction for population structure identification using the Laplacian linear discriminant

Usman Roshan[1,*]

[1]Department of Computer Science, New Jersey Institute of Technology, GITC 4400, University Heights, Newark, NJ 07102, USA

ABSTRACT

Motivation: The identification of population structure from genome-wide SNP data is of significant interest in the population and medical genetics community. A popular solution is to perform unsupervised feature extraction using principal component analysis. Principal component analysis, however, relies only on global properties of the data.

Results: The Laplacian linear discriminant takes into consideration local properties of the data as well and, as we show in this study, it can be extended to the semi-supervised setting. This can then be applied to extract features for identifying population structure when the ancestry of some individuals in some sub-populations of the admixture is known. Using real data we simulate such semi-supervised scenarios and extract features using the Laplacian linear discriminant, kernel principal component analysis, and two recent semi-supervised feature extractors. We show that there is a statistically significant improvement in accuracy when the nearest mean classifier or k-means clustering is applied on the Laplacian linear discriminant features compared to kernel principal component analysis and the other methods.

Availability: All necessary software and data for reproducibility purposes is at http://www.cs.njit.edu/usman/LLDA_pop_structure.

Contact: usman@cs.njit.edu

1 INTRODUCTION

The problem of clustering humans into groups of similar geographical ancestry arises in the fields of medical and population genetics. In medical genetics disease association studies can lead to misleading results if the underlying population structure is not taken into account (Ziv and Burchard, 2003; Marchini et. al., 2004; Xu and Shete, 2005; Devlin and Roeder, 1999). In population genetics this can be used to uncover demographic history and address related scientific questions (Cavalli-Sforza and Feldman, 2003). Consequently several methods have been developed for identifying structure from genome wide single nucleotide polymorphism (SNP) data (Tsai et. al., 2005).

The two prevailing methods are principal component analysis (PCA) (Paschou et. al., 2007), which is an unsupervised feature extraction method, and the model-based Markov Chain Monte Carlo method implemented in STRUCTURE (Pritchard et. al., 2000), where prior ancestry can be specified if available. PCA is very fast and has been shown to separate inter and intra-continental admixtures with high accuracy when followed by the simple k-means clustering algorithm (Paschou et. al., 2007). The model-based approach of STRUCTURE is also considered accurate. However it is very slow and thus prohibitive on admixtures with several thousand SNPs or hundreds of individuals.

In this study consider the following problem. Assume that some individuals in the population have known ancestry. In practice this can be obtained using current benchmarks such as the HAPMAP project (http://www.hapmap.org). How can we then extract features for structure identification while taking the prior ancestry into consideration? We call this semi-supervised feature extraction since the data of individuals with unknown ancestry is also utilized for extracting features. We propose a semi-supervised Laplacian linear discriminant (LLDA) (Tang et. al., 2006) for solving this problem.

LLDA can be considered to be a more general feature extractor than PCA. PCA uses only global properties of the data since it is based on the *total* scatter matrix. LLDA, which is similar to the maximum margin criterion discriminant (Li et. al., 2006), uses both the total scatter matrix, which captures global properties of the data, as well as the *within class* scatter matrix, which captures local properties. Although the within class scatter matrix is normally defined in a supervised framework, it can also be computed in an unsupervised one using Laplacian matrices (Nijima and Okuno, 2007). In this study we extend it to a semi-supervised setting and then extract features with it.

In related work Zhang et. al., 2007, propose a Laplacian based approach that uses must-link and cannot-link constraints for extracting features. Sugiyama et. al., 2007, present a semi-supervised local Fisher discriminant for feature extraction. We contrast our approach with theirs in the Methods Section.

Using real benchmarks we simulate different semi-supervised scenarios and compare our proposed LLDA approach to kernel PCA and the two recent semi-supervised feature extractors. We compare the error of the nearest mean classifier and the k-means clustering algorithm on all the projections. We show that LLDA attains statistically significant higher accuracies than kernel PCA even with a small number of individuals with known ancestry. These accuracies improve as the percent of individuals with known ancestry increases. The two other semi-supervised methods, however, have higher error than our LLDA approach and also kernel PCA on the data considered here.

2 METHODS

Throughout we assume that n represents the total number of individuals in the population and k represents the number of sub-populations in it.

*To whom correspondence should be addressed.

2.1 Background

2.1.1 Encoding the data

We assume that for each individual in the population biallelic SNPs have been assayed. We use the following encoding scheme to convert each individual's set of sequenced SNPs into numerical vectors on which PCA can then be performed. Let g be an individual's set of sequenced SNPs and let g_i represent the i^{th} SNP. g_i can be AA, AB, or BB where A and B are alphabetically ordered SNP bases. Following the encoding of *smartpca* (Patterson et. al., 2006) and similar to the one in (Paschou et. al., 2007) we define the *feature* vector x for g by setting x_i to 0 if g_i is AA, 1 if AB, and 2 if BB. Let each such vector x for the i^{th} individual be the i^{th} column of the data matrix X, i.e. $X = [x_1, \ldots, x_n]$.

2.1.2 Principal component analysis (PCA)

Let x be a random variable that represents the feature vector of an individual's SNP genotype as defined above. Suppose we are interested in computing a projection of x onto one dimension such that the variance (i.e. spread) of the projected data is maximized. In other words we want to find w such that $Variance(w^T x)$ is maximized subject to $w^T w=1$. This yields the optimization problem

$$\max_w w^T \Sigma^T w \qquad \text{subject to } w^T w = 1 \tag{1}$$

where $\Sigma=E((x-\mu)(x-\mu)^T)$ is the covariance of x and $\mu=E(x)$ is the expected value of x. Using Lagrange multipliers one can show that the eigenvector of Σ with the largest eigenvalue is the solution to problem (1). The eigenvector with the next largest eigenvalue yields a projection orthogonal to the first one and of maximum variance (Alpaydin, 2004).

In practice we use the sample covariance matrix instead of Σ. Given the data matrix $X = [x_1, \ldots, x_n]$ (as described in the encoding) we define $X' = [x_1-m, \ldots, x_n-m]$, where x_i is the feature vector of the i^{th} individual and $m=\Sigma x_i/n$ is the mean feature vector of the population. The sample covariance matrix is then defined as $X'X'^T$. This is also called total scatter matrix and is equivalent to

$$S_t = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)(x_i - m)^T$$

The eigenvector of S_t with the largest eigenvalue (also called the first principal component vector) gives the projection of maximum variance in the sample. The eigenvector with the next largest eigenvalue yields a projection orthogonal to the first one and also of maximum variance. In unsupervised scenarios PCA can be very helpful in elucidating clusters in the data.

2.1.3 Maximum margin discriminant analysis (MMC)

In a supervised scenario one can use the standard Fisher discriminant (Alpaydin, 2004). Another supervised method is the recent maximum margin discriminant (MMC) (Li et. al., 2006) that does not suffer from singularity problems like its Fisher counterpart. Define the within class scatter matrix S_i for class i as

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_j^{(i)} - m^{(i)})(x_j^{(i)} - m^{(i)})^T$$

the overall within class scatter matrix as

$$S_w = \frac{1}{n}\sum_{k=1}^{c} n_k S_k = \frac{1}{n}\sum_{k=1}^{c}\sum_{j=1}^{n_k}(x_j^{(k)} - m^{(k)})(x_j^{(k)} - m^{(k)})^T$$

and the between class scatter matrix as

$$S_b = \frac{1}{n}\sum_{k=1}^{c} n_k (m^{(k)} - m)(m^{(k)} - m)^T$$

where $x_j^{(i)}$ is the j^{th} feature vector of class i, c is the number of classes, n_i is the size of class i, $m^{(i)}$ is the mean feature vector of class i, and m is the mean feature vector of the entire dataset. The maximum margin criterion for feature extraction is defined as (Li et. al., 2006)

$$J = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{c} p_i p_j d(C_i, C_j)$$

where p_i and p_j are class prior probabilities and $d(C_i,C_j)$ is the interclass distance. Define the interclass distance $d(C_i,C_j)$ as $d(C_i,C_j)=d(m_i,m_j)-tr(S_i)-tr(S_j)$ where $d(m_i,m_j)$ is the Euclidean distance between the mean vectors m_i and m_j and $tr(S_i)$ is the overall variance of class S_i. Then if $p_i=n_i/n$ and $d(C_i,C_j)=d(m_i,m_j)-tr(S_i)-tr(S_j)$, it can be shown that $J=tr(S_b-S_w)$ (Li et. al., 2006). The linear MMC discriminant aims to find a matrix $W = [w_1, w_2, \ldots, w_d]$ that maximizes

$$tr(W^T(S_b - S_w)W)$$

subject to $(w_k^T w_k=1)$ for all $k \in [1,d]$. This is equivalent to solving

$$\max_{w_1 \ldots w_d} \sum_{k=1}^{d} w_k^T (S_b - S_w) w_k$$
$$\text{subject to } w_k^T w_k = 1 \text{ for } k = 1, \ldots, d$$

Using Lagrange multipliers it can be shown that the d largest eigenvector of S_b-S_w are the solution to W. The projection of the data matrix X is then given by $W^T X$. Since $S_t=S_w+S_b$, S_b-S_w can be rewritten as S_t-2S_w (Nijima and Okuno, 2007). Thus MMC also takes into consideration local properties of the data (by considering S_w) whereas PCA only considers S_t.

2.1.4 Laplacian linear discriminant analysis (LLDA)

In order to describe this discriminant we first define the Laplacian matrix of a weighted graph. Suppose we are given a weighted graph G with n nodes and its associated weight matrix $W = \{w_{ij} : i,j \in [1,n]\}$. Then the Laplacian L of G is defined as (Tang et. al., 2006)

$$L[i,j] = \begin{cases} -w_{ij} & i \neq j \\ d_i = \sum_{k=1}^{n} w_{ik} & i = j \end{cases} \tag{2}$$

Now we describe LLDA. First note that the matrices S_t and S_w defined earlier can also be written as (Nijima and Okuno, 2007)

$$S_t = \frac{1}{n}X(I - \frac{1}{n}ee^T)X^T = \frac{1}{n}X(I - W_g)X^T = \frac{1}{n}XL_gX^T \tag{3}$$

$$S_w = \frac{1}{n}X(I - \sum_{k=1}^{c}\frac{1}{n_k}e^{(k)}e^{(k)T})X^T = \frac{1}{n}X(I - W_l)X^T = \frac{1}{n}XL_lX^T$$

where e is n dimensional with all entries set to 1 and the i^{th} entry of $e^{(k)}$ is set to 1 if x_i belongs to class k and 0 otherwise. I-W_g can be viewed as the Laplacian L_g of a *global* graph where all vertices are connected and each edge has weight $1/n$. Similarly I-W_l is the Laplacian L_l of a *local* graph such that all vertices belonging to the same class are connected with weight $1/n_k$.

In terms of the graph Laplacians we can represent S_t-$2S_w$ as $(1/n)X^T(L_g$-$2L_l)X$ using (3). MMC represented in this form is known as Laplacian linear discriminant analysis (Nijima and Okuno, 2007). The d leading eigenvectors of $(1/n)X^T(L_g$-$2L_l)X$ form the columns of W and the projection is then given by W^TX. Note that so far the Laplacian of the local graph is well-defined only under a supervised scenario. However, it can also be used in an unsupervised setting (Nijima and Okuno, 2007) and a semi-supervised one as we show below.

2.1.5 Related work

Zhang et. al., 2007, also use a Laplacian graph based approach for semi supervised dimensionality reduction. They specify must-link and cannot-link pairwise constraints in the weight matrix and then proceed to compute its Laplacian. The largest eigenvectors of the Laplacian are then used to project the data. Our approach, however, is based on the LLDA, which in turn is based on the MMC criterion. The LLDA approach takes both the global and local properties into consideration in separate Laplacians. Furthermore, the Laplacians in our case are constructed from nearest neighbor graphs that also incorporate prior information (see Subsection 2.2.1). The approach of Zhang et. al., 2007, on the other hand, does not use nearest neighbor graphs for construction of the weight matrix.

Sugiyama et. al., 2007, propose a semi-supervised local Fisher discriminant for dimensionality reduction. They use regularized between-class and within-class scatter matrices and define a trade-off parameter to control the regularization. Their approach is based on the local Fisher discriminant whereas LLDA (and subsequently our method) is closer to the MMC criterion. Li et. al., 2006, have shown MMC to produce lower error than the Fisher discriminant on facial recognition benchmarks.

2.2 Our contribution

2.2.1 Semi-supervised Laplacian linear discriminant

In this study we describe a semi-supervised extension. Suppose that the ancestry of some individuals from some sub-populations in the admixture is known. In order to take this into account for feature extraction in an LLDA framework we proceed as follows. First we define the weight matrices for the local and global graphs as follows. Our definitions of weight matrices are similar to the ones in Nijima and Okuno, 2007, except that we incorporate prior knowledge.

$$W_l[i,j] = \begin{Bmatrix} 100\,p_{ij} & \text{if ancestry of } i \text{ and } j \text{ are specified and } cl(i) = cl(j) \\ k(i,j) & \text{if } i \text{ is among } m \text{ nearest neighbors of } j \text{ or vice - versa} \\ & 0 \text{ otherwise} \end{Bmatrix}$$

$$W_g[i,j] = \begin{Bmatrix} k(i,j) & \text{if } i \neq j \\ 0 & \text{if } i = j \end{Bmatrix}$$

where $k(i,j)$ is the similarity of individuals i and j, and $cl(i)$ returns the class (ancestry) of individual i. p_{ij} is defined as

$$p_{ij} = \sum_k q_{ik} q_{jk}$$

where q_{ik} is the probability that individual i belongs to sub-population k of the admixture.

Current benchmarks are obtained from individuals where both parents and grandparents of an individual are from the same ancestry (e.g. http://www.hapmap.org). Thus q_{ik} is usually 1 or 0. However if an individual has different ancestry from the mother and father then q_{ik} will be a probability between 0 and 1.

We use prior knowledge also in the construction of the nearest neighbor graph. For a given individual i whose ancestry is specified we calculate its m nearest neighbors as follows. Let C be the set of individuals with the same specified ancestry as that of i. We consider all individuals in C to be closer to i even if they are considered far under a given distance metric. Thus, we add them to the nearest neighbors of i sorted by distance to i. If $|C| < m$ then we sort the remaining individuals, i.e. the total population without C, by their distance to i and add the closest m-$|C|$ ones to the nearest neighbor set of i. If the prior ancestry of i is not specified we calculate its distance to all other individuals in the admixture and include the m nearest ones in the nearest neighbor set of i.

After computing the matrices W_g and W_l we calculate the global and local Laplacians according to (2). We then form the matrix $(1/n)X^T(L_g$-$2L_l)X$ and compute the d leading eigenvectors as the solution to W, where d is the desired number of reduced dimensions.

Note that the semi-supervised LLDA as well as the unsupervised one in Nijima and Okuno, 2007 both include MMC as a special case. If we set $W_g[i,j]=1/n$ for all i,j, and $W_l[i,j]=1/n_k$ if i and j are both in the same class k and 0 otherwise, we then obtain MMC.

2.2.2 Our implementation: Kernel-PCA + LLDA

The dimensions of $(1/n)X^T(L_g$-$2L_l)X$ can be very large when thousands of SNPs are given. This can considerably slow down the eigenvector computations. To overcome this we follow the spirit of PCA+LDA algorithms for image recognition (Yang et. al., 2005). We first compute the full PCA projection of the SNP data. This can be done efficiently using kernel PCA (Scholkopf and Smola, 2002). Kernel PCA also allows us to model non-linear relationships in the data. After the kernel PCA transformation each individual is now represented by an n dimensional vector (where n is the number of individuals). This is a significant reduction in dimension. SNPs can be in the order of hundreds of thousands or millions whereas the number of individuals in the population may be much smaller.

We treat the kernel PCA projection as the new data matrix, i.e. column i of the data matrix X is replaced with the kernel PCA projected vector of the i^{th} individual. We then apply LLDA on this matrix. Although our method is technically kernel-PCA + semi-supervised LLDA we refer it to as LLDA hereon.

All software and the datasets (including the kernel PCA projections) are available from http://www.cs.njit.edu/usman/LLDA_pop_structure. In particular, scripts and data required to reproduce the results shown in this study are also provided.

2.2.3 Parameters

In the computation of W_l we set $k(i,j)$ to 1 everywhere, although note that Euclidean, dot product, or Gaussian kernels may also be used (Nijima and Okuno, 2007). For the nearest neighbor search we use the Euclidean distance between the kernel-PCA projected vectors. The number of dimensions in the PCA projected vectors used for the Euclidean distance calcula-

tion is set to 10 and the parameter m (in the definition of W_i) is also set to 10. We experimented with fewer dimensions and smaller values of m but observed negligible differences in error.

Since we use benchmarks where both maternal and paternal parents and grandparents have the same ancestry, q_{ik} is set to 1 for the appropriate sub-population k and 0 for remaining ones if i^{th}'s prior ancestry is available. Thus p_{ij} is always 1 when both i and j have the same ancestry.

We used the kernel PCA program gist-kpca of the GIST software suite (http://bioinformatics.ubc.ca/gist) for computing kernel-PCA projections with the polynomial degree two kernel. We experimented with Gaussian and higher order polynomials but did not observe a significant difference in error. We implemented LLDA using Perl and employed the Perl Math Cephes package (http://search.cpan.org/dist/Math-Cephes) for matrix operations.

3 RESULTS

In order to study the performance of our proposed approach we simulate two different semi-supervised scenarios. In the first scenario we assume that for a given admixture some individuals from *each* sub-population have known ancestry. In the second one some individuals from only *some* sub-populations have known ancestry. Note that the semi-supervised scenario does not affect the PCA projection since that is an unsupervised method. However, it produces different LLDA or other semi-supervised projections.

In order to compare different projections under the first scenario we apply the nearest mean classifier (Alpaydin, 2004). Since training samples are available from each sub-population in this case it is straightforward to apply this classifier. The procedure is simple: first calculate the training means of each sub-population and then assign each individual to the sub-population with the shortest Euclidean distance to its mean. In the second scenario we use the k-means clustering algorithm (Alpaydin, 2004) on both the projections. In this situation only some individuals of some sub-populations have known ancestry and therefore it is not possible to apply a classifier.

When applying the nearest means or k-means methods on a projection we consider only the leading k dimensions, where k is the number of sub-populations in the admixture. Note that this number is not required to compute the LLDA projection. We only use it in order to compare the projections under the nearest means and k-means algorithms. In practice the number k can be estimated using various methods (Sanguinetti et. al., 2005).

Given a true and estimated classification we use the classification error rate for determining its error. Let $T[i] \in [1,k]$ be the true sub-population ID of individual i and similarly let $C[i]$ be the estimated sub-population ID. The error rate is then

$$\left(100 \times \left|\{j : T[j] \neq C[j], j \in [1,n]\}\right|\right) / n$$

If a clustering is given then the above approach is not applicable. Instead we first have to make sure that the estimated sub-population IDs match the true ones by the following method: for each estimated sub-population we assign it the ID that is maximum among the true population IDs of all individuals in it. We then proceed with the same formula. We measure the statistical significance of the difference in errors between two methods using the Wilcoxon signed rank test (Kanji, 1999).

As shown in earlier studies PCA followed by k-means can separate sub-populations of the HAPMAP dataset (such as Chinese and Japanese individuals) with high accuracy (Paschou et. al., 2007). We focus here on the recent large dataset from Noah Rosenberg's lab (Jakobsson et. al., 2008). This contains 525,9210 genome-wide SNPs of 485 individuals from 29 populations. From this large dataset we extract individuals from three different regions as specified in the dataset:

- **East Asia**: 10 from Cambodia, 15 from Siberia, 49 from China, and 16 from Japan; 459,188 SNPs without missing entries
- **Africa**: 32 Biaka Pygmy individuals, 15 Mbuti Pygmy, 24 Mandenka, 25 Yoruba, 7 San from Namibia, 8 Bantu of South Africa, and 12 Bantu of Kenya; 454,732 SNPs without missing entries.
- **Middle East**: 43 Druze from Israel-Carmel, 47 Bedouins from Israel-Negev, 26 Palestinians from Israel-Central, and 30 Mozabite from Algeria-Mzab; 438,596 SNPs without missing entries.

We conducted several preliminary studies to determine which methods to present in the main results. We first compared kernel-PCA to the implementation of the smartpca program (Patterson et. al., 2006) and found it to have lower error when evaluated under k-means.

We also compared the SSDR method of Zhang et. al., 2007, and the semi-supervised local Fisher discriminant (SSLFDA) of Sugiyama et. al., 2007, to LLDA under the first semi-supervised scenario. We implemented both of these methods in Perl and provide them in our software distribution at http://www.cs.njit.edu/usman/LLDA_pop_structure. In line with our semi-supervised LLDA implementation (kernel-PCA + LLDA) we apply both SSDR and SSLFDA on the full kernel PCA projection. We found both SSDR and SSLFDA to have higher errors than LLDA and even the original kernel PCA on the data considered here. Subsequently we omit them from the remaining results.

3.1 Random number of individuals with known ancestry from each sub-population of the admixture

In the first semi-supervised scenario we assume that some individuals from each sub-population of the admixture have known ancestry. In order to apply LLDA here we would need a minimum of two individuals from each sub-population with known ancestry.

3.1.1 Between 2 and 4 random individuals of each sub-population with known ancestry (for Middle Eastern admixture between 2 and 8)

We simulate datasets with known ancestry as follows. Assume that the admixture has p sub-populations. For each sub-population $j \in [1,p]$ we generate a random $r_j \in [2,4]$ and randomly select r_j individuals from sub-population j to have known ancestry. For the Middle Eastern admixture, however, we use the range [2,8]. We generate 200 such random datasets. On each such dataset we extract LLDA features. The kernel-PCA features are unsupervised and so independent of the prior ancestry.

The number of random individuals with known ancestry is much smaller than the total in the admixture. For the datasets in this subsection the mean number of individuals with known ancestries in the East Asian, African, and Middle Eastern admixtures are 11.0, 19.1, and 18.3 respectively. These constitute 12.2%, 15.5%, and 12.5% of the three admixtures respectively.

Since each sub-population has a minimum of two individuals with known ancestry we can apply the simple nearest means classifier on the two projections and measure their error. In Table 1 we report the mean error on both the projections (averaged over the 200 random trials). On all three admixtures the LLDA error is lower by a statistically significant margin.

Table 1. Error of nearest means classifier. * denotes Wilcoxon signed rank test p-value < 0.05.

Admixture	Kernel PCA	Laplacian linear discriminant
East Asia	5.3	4.3*
Africa	9.1	5.3*
Middle East	23.8	19.4*

3.1.2 Between 2 and 20% random individuals of each sub-population with known ancestry

We simulate 200 random datasets with known ancestry in a manner similar to the one described in the previous section. However, the range for each r_j is $[2,0.2n_j]$ where n_j is the number of individuals in sub-population j of the admixture. The mean percent of individuals with known ancestry is 13.5%, 15.1%, and 12% for the three admixtures respectively. LLDA still performs significantly better as Table 2 shows.

Table 2. Caption as Table 1.

Admixture	Kernel PCA	Laplacian linear discriminant
East Asia	4.7	3.7*
Africa	8.6	4.3*
Middle East	23.2	18*

3.2 Random number of individuals with known ancestry but only from some sub-populations of the admixture

In the second semi-supervised scenario we assume that only some individuals from some sub-populations have known ancestry. This presents a more difficult scenario since now a classification method cannot be applied. In order to compare the two projections in this scenario we apply the popular k-means clustering method. This may not necessarily the best way of comparison but it still gives us some idea of how well separated the two projections are. We run k-means 100 times on each input and report the error of the clustering with the highest objective function value.

3.2.1 Between 2 and 4 random individuals of some sub-populations with known ancestry (for Middle Eastern admixture between 2 and 8)

Assume that the admixture has p sub-populations. We first select a random number $p' \in [1,p]$ and then pick p' random sub-population IDs (from 1 through p) using Fisher-Yates sampling (Knuth, 1998). For each selected sub-population j we generate a random $r_j \in [2,4]$ and randomly select r_j individuals from sub-population j to have known ancestry (as before). Again for the Middle Eastern admixture we use the range [2,8]. We generate 200 such random datasets. On each such dataset we extract LLDA features. The kernel-PCA features are unsupervised and so independent of the prior ancestry.

The number of selected sub-populations p' and number of individuals per selected sub-population varies across these datasets. For example for the 200 datasets in this subsection the mean values of p' for the East Asian, African, and Middle Eastern admixtures are 2.6, 3.9, and 2.5. These values are very similar for the data in the proceeding subsections.

The number of individuals with known ancestry is also much smaller than the total admixture sizes. For the data in this subsection the mean number of individuals with known ancestry are 7.1, 10.8, and 11.8 for the East Asian, African, and Middle Eastern admixtures. These constitute 7.8%, 8.8%, and 8.1% of the respective admixtures.

We apply the k-means clustering algorithm on both the kernel-PCA and LLDA projections. K-means on the kernel-PCA projection has the same error across the different datasets because the prior ancestry does not affect the projection. In Table 3 we report the k-means error averaged over the 200 random trials. Except for the East Asian admixture LLDA has statistically significant lower errors.

Table 3. Error of k-means. * denotes Wilcoxon signed rank test p-value < 0.05.

Admixture	Kernel PCA	Laplacian linear discriminant
East Asia	8.9	7.1
Africa	16.3	14.4*
Middle East	29.5	23.0*

3.2.2 Between 2 and 20% random individuals of some sub-populations with known ancestry

We proceed as in the previous subsection. We first select *p'* random population IDs. However, this time we select r_j randomly from $[2,0.2n_j]$ where n_j is the number of individuals in the selected sub-population *j* of the admixture. The mean percent of individuals with known ancestry is 9%, 9.1%, and 7.4% for the three admixtures respectively. As Table 4 shows LLDA performs significantly better than kernel-PCA.

Table 4. Caption as Table 3.

Admixture	Kernel PCA	Laplacian linear discriminant
East Asia	8.9	5.9*
Africa	16.3	13.7*
Middle East	29.5	22.7*

3.2.3 Between 2 and 35% random individuals of some sub-populations with known ancestry

We proceed as previously but select r_j randomly from $[2,0.35n_j]$. The mean percent of individuals with known ancestry is 12.2%, 11.9%, and 12% for the three admixtures respectively. Table 5 shows the improvements gained from LLDA.

Table 5. Caption as Table 3.

Admixture	Kernel PCA	Laplacian linear discriminant
East Asia	8.9	5.5*
Africa	16.3	14.0*
Middle East	29.5	21.8*

3.2.4 Between 2 and 50% random individuals of some sub-populations with known ancestry

Finally we select r_j randomly from $[2,0.5n_j]$. The mean number of individuals with known ancestry (over the 200 trials) is still small for the data in this subsection: 15.5 for East Asian admixture, 19.3 for African, and 24.6 for the Middle Eastern. These constitute 17.2%, 15.7%, and 16.8% of the three admixtures respectively. We summarize the results in Table 6.

Table 6. Caption as Table 3.

Admixture	Kernel PCA	Laplacian linear discriminant
East Asia	8.9	4.5*
Africa	16.3	12.2*
Middle East	29.5	19.8*

3.3 Discussion

On the HAPMAP dataset PCA followed by k-means separates the Japanese and Chinese sub-populations with 1% error (Paschou et. al., 2007). We observed the same error with kernel PCA. On the admixtures considered here however, PCA error is considerably higher. For example, on the Middle Eastern admixture kernel PCA followed by k-means reaches an error of 29.5%. This admixture is particularly hard because of the closely related Bedouin and Palestinian sub-populations. LLDA gives an improvement of almost 10% when mean percent of known ancestry is 16.8% over randomly selected sub-populations of this admixture (see Table 6).

The percent of individuals with known ancestry is intentionally kept small throughout our experiments in order to simulate hard scenarios. In practice prior ancestry of many individuals may be available which in turn will provide a greater advantage with LLDA. This trend can be observed from Tables 4 through 6: as the mean number of individuals with known ancestry increases the LLDA error drops. Our LLDA implementation is also very fast: for any of the three given admixtures it finishes in a few seconds.

4 CONCLUSIONS

We proposed a semi-supervised Laplacian linear discriminant for extracting features for identifying population structure when the ancestry of some individuals is known in advance. Using real benchmarks we simulate various semi-supervised scenarios and show that LLDA outperforms kernel PCA by a statistically significant margin. In comparison to two recent semi-supervised feature extractors, LLDA and kernel PCA have lower error on the data considered here. The proposed LLDA method is fast, can be easily implemented, and accommodates mixed prior ancestries.

ACKNOWLEDGEMENTS

Computational experiments were performed on the CIPRES cluster supported by NSF award 033-1654. We thank system administrators at New Jersey Institute of Technology and the San Diego Supercomputing Center for their support.

REFERENCES

Alpaydin, E. (2004) Introduction to Machine Learning, MIT Press

Cavalli-Sforza L., Feldman M. (2003) The application of molecular genetic approaches to the study of human evolution, *Nature Genetics* 33: 266–275.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997–1004

Jakobsson M., Scholz S.W., Scheet P., Gibbs J.R., VanLiere J.M., Fung H-C., Szpiech Z.A., Degnan J.H., Wang K., Guerreiro R., Bras J.M., Schymick J.C., Hernandez D.G., Traynor B.J., Simon-Sanchez J., Matarin M., Britton A., van de Leemput J., Rafferty I., Bucan M., Cann H.M., Hardy J.A., Rosenberg N.A., Singleton A.B. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.

Kanji G. K. (1999) 100 Statistical Tests, London, U.K., Sage Publications

Knuth D. E. (1998) The Art of Computer Programming volume 2, 3rd edition, 145-146

Li H., Tiang J., Zhang J. (2006) Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks* 17(1): 157-165

Marchini J, Cardon L, Phillips M, Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nature Genetics* 36: 512–517.

Nijima S., Okuno Y. (2007) Laplacian linear discriminant analysis approach to unsupervised feature selection, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* PP(99):1-1, doi:10.1109/TCBB.2007.70257

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, et al. (2007) PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. *PLoS Genetics* 3(9): e160 doi:10.1371/journal.pgen.0030160

Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2: e190. doi:10.1371/journal.pgen.0020190

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Sanguinetti G., Laidler J., Lawrence N. D. (2005) Automatic determination of the number of clusters using spectral algorithms, *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, 55-60

Scholkopf, B., & Smola, A. (2002). Learning with kernels. Cambridge, MA, MIT Press.

Sugiyama M., Ide T., Nakajimi S., Sese J. (2007) Semi-supervised local Fisher discriminant analysis for dimensionality reduction, *Proceedings of the Workshop on Information-Based Induction Sciences*, Tokyo, Japan

Tang H, Fang T., Shi P-F. (2006) Laplacian linear discriminant analysis, *Pattern Recognition*, 39: 136-139

Tsai H, Choudhry S, Naqvi M, Rodriguez-Cintron W, Burchard E, et al. (2005) Comparison of three methods to estimate genetic ancestry and control for stratification in genetic association studies among admixed populations, *Hum Genet* 118: 424–433.

Xu H., Shete S. (2005) Effects of population structure on genetic association studies. *BMC Genetics* 6(Suppl 1): S109

Yang J., Frangi A.F.,Yang J-Y., Zhang D., Zhong J. (2005) KPCA plus LDA: a complete kernel Fisher discriminant framework for feature extraction and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(2): 230-244

Zhang D., Zhou Z-H., Chen S., (2007) Semi-supervised dimensionality reduction, *Proceedings of the 2007 SIAM International Conference on Data Mining*, Minneapolis, Minnesota, USA

Ziv E, Burchard E (2003) Human population structure and genetic association studies. *Pharmacogenomics* 4: 431-441.