Article

A Comparison Study on Nonlinear Dimension Reduction Methods with Kernel Variations: Visualization, Optimization and Classification

Katherine C. Kempfert[1,†], Yishi Wang [2*], Cuixian Chen [2], and Samuel W.K. Wong [3]

[1] University of Florida; kkempfert2@ufl.edu
[2] University of North Carolina Wilmington; {wangy, chenc}@uncw.edu
[3] University of Waterloo; samuel.wong@uwaterloo.ca
* Correspondence: wangy@uncw.edu; Tel.: +1-910-962-3292
† Current address: The University of Florida. Gainesville, FL 32611

Version May 23, 2019 submitted to Intelligent Data Analysis

Abstract: Because of high dimensionality, correlation among covariates, and noise contained in data, dimension reduction (DR) techniques are often employed to the application of machine learning algorithms. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and their kernel variants (KPCA, KLDA) are among the most popular DR methods. Recently, Supervised Kernel Principal Component Analysis (SKPCA) has been shown as another successful alternative. In this paper, brief reviews of these popular techniques are presented first. We then conduct a comparative performance study based on three simulated datasets, after which the performance of the techniques are evaluated through application to a pattern recognition problem in face image analysis. The gender classification problem is considered on MORPH-II and FG-NET, two popular longitudinal face aging databases. Several feature extraction methods are used, including biologically-inspired features (BIF), local binary patterns (LBP), histogram of oriented gradients (HOG), and the Active Appearance Model (AAM). After applications of DR methods, a linear support vector machine (SVM) is deployed with gender classification accuracy rates exceeding 95% on MORPH-II, competitive with benchmark results. A parallel computational approach is also proposed, attaining faster processing speeds and similar recognition rates on MORPH-II. Our computational approach can be applied to practical gender classification systems and generalized to other face analysis tasks, such as race classification and age prediction.

Keywords: Dimension Reduction; PCA; LDA; FDA; KPCA; KFDA; SKPCA; SVM; Parameter Optimization; Gender Classification; MORPH-II.

1. Introduction

Due to advances in data collection and storage capabilities, the demand has been growing substantially for gaining insights into high-dimensional, complex-structured, and noisy data. Researchers from diverse areas have applied DR techniques to visualize and analyze such data [1,2]. DR techniques are also helpful to address the issues of collinearity and "$p \gg n$" (i.e., number of features exceeding the sample size in a dataset), by projecting the data into a lower dimensional space with less correlation, so that classical statistical methods can be applied [3]. Principal Component Analysis (PCA) [4,5] is a well-studied algorithm with the goal of projecting input features onto a lower dimensional subspace while preserving the largest variance possible; lower dimensionality permits easier visualization, for example via heat maps. While PCA is a fully automatic algorithm, DR techniques that account for domain expertise via user input have also been more recently studied [6,7]. For classification problems, in which the label information as the response variable is available, Linear Discriminant Analysis (LDA) (sometimes referred to as Fisher's Discriminant Analysis (FDA)) can be used for DR by minimizing intra-class variation and maximizing inter-class variation [8,9]. Since PCA only utilizes the correlation or covariance matrices, it is considered an unsupervised approach,

whereas LDA is considered a supervised approach with labeling information built into its objective function. Despite the dissimilarities, both PCA and LDA search for linear combinations of the features and, therefore, can be applied in linearly separable types of problems [10]. The main challenge is that many problems in practical applications of machine learning are nonlinear [11,12]. For nonlinear DR, kernel methods are popular choices because of their flexibility [13–15], e.g, Kernel PCA [16], Kernel LDA for two classes [17], and more generalized Kernel LDA for multiple classes [18]. For kernel methods, it is also possible to design specialized kernels based on domain knowledge of a problem [19,20].

Given the problems in image analysis of high dimensionality and complex correlation structures, DR techniques are often a necessary step [21]. Thus, variants of PCA, LDA, and their kernel extensions have been popular in computer vision with applications of image classification and discrimination [22–24]. Studies include Eigenfaces [25], Fisherfaces [26], face recognition with KPCA [27], face recognition with Kernel Direct LDA [28], 2D-PCA [29], 2D-LDA [30], among many others. When there are sufficient labeled face images, LDA is experimentally reported to outperform PCA for face recognition [26]. In the case of a small training set, the conclusion could be reversed [23]. Studies comparing classification performance of PCA, LDA, and their kernel variations include [31,32]. The connections among KLDA, KPCA, and LDA are further discussed in [33]. By incorporating labeling information into the construction of the objective function, Supervised Kernel PCA (SKPCA) [34] has been proposed for visualization, regression, and classification. A modified version of SKPCA for classification problems can be found in [35]. These studies suggest that SKPCA works well in practice among different DR algorithms [36–38]. Moreover, it has been found in [39] that with bounded kernels, projections from SKPCA are uniformly converging, regardless of the input features' dimension.

2. Associated Work

In recent years, facial demographic analysis has become popular in computer vision, because of its broad applications in human-computer interaction (HCI), security, surveillance, and marketing, which can benefit from the automatic estimation of characteristics like age, gender, and race. Recent surveys on demographic estimation from biometrics are presented in [40,41]. Specifically, a major task is gender classification, aiming to automatically determine if a person is male or female. Beyond computer vision, the topic has been studied extensively by anthropologists, sociologists, and psychologists. Gender can easily be identified by humans, achieving 96% accuracy in an experiment classifying photographs of adult faces [42]. Automating gender classification has been a priority in real-world applications. A number of biometrics have been used to identify gender, including face, voice, gait, handwriting, and even the iris [41]. However, gender classification from faces is the most common, probably because photography of faces is non-intrusive and ubiquitous. Ng et al. provide a survey of gender classification via face and gait [43].

Gender classification with faces launched in 1990, when neural networks were applied directly to pixels from face photographs [44,45]. Many other early studies utilized the *geometric-based* approach to represent human faces, relying on measurements of facial landmarks [46,47]. Though intuitive, such approaches are sensitive to the placement of landmarks, which can only accommodate frontal representations of the face, and may omit some important information from the face (such as texture of the skin). In recent years, the *appearance-based* methods have been more commonly adopted, which rely on a transformation of an image's pixels [48–50]. Such methods capture both the geometric relationships of the face and texture information. However, a drawback is their sensitivity to illumination and viewpoint variations. Other issues are associated with the high dimensionality of the transformed pixels, which will be discussed further in the next paragraph. Some most recent gender classification studies involve convolutional neural networks (CNN) [51–54]. Though CNNs have reached state-of-the-art accuracy rates, they are known to be less interpretable than some other methods.

Pixels often contain high redundancy and noise, which cannot be removed completely by pre-processing steps. Hence, the vectors resulting from *appearance-based* feature extraction methods genetically inherit redundancy and noise. Popular image feature extraction methods include local texture techniques such as local binary patterns (LBP) [55–58], Gabor filters [59], biologically-inspired features (BIF) [60,61], and histogram of oriented gradients (HOG) [60]. Such methods could lead to a high dimension of extracted features, thwarting practical applications by increasing runtime and memory consumption. When "$p \gg n$", for which the dimension of the feature space exceeds the sample size of the dataset, a fundamental assumption of many standard statistical procedures is violated. Additionally, collinearity of features can cause numerical problems, while noisy features can obscure true relationships with the response variable and hinder predictive performance. These significant issues motivate the use of DR techniques. The fundamental goal of DR is to extract and retain information in a lower dimensional space. Many of these methods fall under manifold learning, identifying a low-dimensional manifold embedded in a high-dimensional ambient space [62].

Even though PCA and LDA have been widely considered as popular and effective approaches for DR in machine learning, their kernel versions are much less investigated. To our best knowledge, KPCA, KLDA, and SKPCA have never before been directly compared on visualization and classification performance through simulations and practical applications to face image analysis problems.

Our main contributions in this study can be summarized as follows. (1) The nonlinear manifold learning projections for KPCA, KLDA, and SKPCA are directly compared with visualization through simulated datasets. (2) Motivated by the nonlinear nature of soft-biometric analysis problems, we utilize KPCA, KLDA, and SKPCA for dimension reduction on four types of appearance-based extracted features (BIF, HOG, LBP, and AAM) for the gender classification task. Moreover, the classification performance is compared systematically on parameter optimization. (3) For applications to practical large-scale systems, we propose an additional parallel computational framework that can decrease runtime while maintaining similar classification rates.

The remainder of the paper is structured as follows. In Section 3, we review the theory of KPCA, SKPCA, and KLDA. In Section 4, we conduct simulation studies to visualize projections. We propose our machine learning methods for gender classification on Morph-II in Section 5. The comparative performance of KPCA, SKPCA, and KLDA on Morph-II is presented and discussed in Section 6. The performance of these DR methods is further compared in Section 7 through application to the FG-NET dataset. The computational framework for large-scale practical systems is proposed in Section 8 and investigated on Morph-II. Finally, we conclude and offer future directions of research in Section 9.

3. Kernel-Based Dimension Reduction Methods

The nonlinearity in a classification problem can often be addressed by kernel-based DR methods, with the appropriate choice of kernels. The driving reasons are the nonlinearity of chosen kernels, flexibility of tuning parameter selection, and most importantly, the kernel tricks. Mercer's theorem guarantees that a symmetric positive-definite function can be written as the sum of a convergent sequence of product functions, which potentially project the data into infinite-dimensional space [63]. Thus, it is feasible to separate the data in the new space. On the other hand, Representer Theorem shows that the solution for certain kernel methods lies in the finite-dimensional span of the training data [63,64]. This is very helpful, since we do not need to compute the coordinates of the projected data in the infinite-dimensional space, but only the inner products between all pairs of data in the feature space.

3.1. Notations

With the goal of emphasizing the connections between KPCA, SKPCA, and KLDA, we define the following notations for classification problems.

Let $\mathcal{X}$ be the feature space, a non-empty subset in $\mathbb{R}^p$ with p as the number of covariates for each subject. Let $\mathcal{Y}$ be the space for the response variable, a subset in $\mathbb{R}$. Let $\{(x_1, y_1), \cdots, (x_n, y_n)\} \subset$

$\mathcal{X} \times \mathcal{Y}$ be a series of n independent observations following a joint probability measure $P_{\mathcal{X},\mathcal{Y}}$. Let $Y = [y_1, y_2, \cdots, y_n]^T$ denote the outcomes of the response variable. Let X be an $n \times p$ feature matrix, with x_i^T as the i-th row for $i = 1, \cdots, n$, and $x^{(l)} \in \mathbb{R}^n$ for $l = 1, \cdots, p$ as its l-th column. Thus, the X matrix can be written as:

$$X = \left[x_1, x_2, \cdots, x_n\right]^T = \left[x^{(1)}, x^{(2)}, \cdots, x^{(p)}\right].$$

Without loss of generality, we may assume that each column of the X matrix is normalized, such that the mean of $x^{(l)}$ is 0 and standard deviation is 1, for $l = 1, \cdots, p$.

Let Σ be the sample covariance matrix of X. We then have

$$\underset{p\times p}{\Sigma} = \frac{1}{n-1} X^T X = \frac{1}{n-1} \sum_{i=1}^{n} x_i {x_i}^T. \tag{1}$$

Let $\mathcal{F}$ be a reproducing kernel Hilbert space on $\mathcal{X}$ from a kernel function $k(\cdot, \cdot)$, which is a Mercer kernel (symmetric and positive-definite), and $\mathcal{G}$ be a reproducing kernel Hilbert space on $\mathcal{Y}$ from a kernel function $l(\cdot, \cdot)$.

For the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, its associated space $\mathcal{F}$ may be infinite-dimensional, but with some additional conditions, the minimizer of a regularized risk function lies in the finite span of the training observations [63]. Additionally, it has been shown [63] that there exists a function

$$\phi : \mathcal{X} \to \mathcal{F} \tag{2}$$

such that for all $x, x' \in \mathcal{X}$,

$$k(x, x') = < \phi(x), \phi(x') >, \tag{3}$$

where $< \cdot >$ is the dot product. Let K be a matrix such that its ij-th element is $k(x_i, x_j)$. We then have

$$K = \{k(x_i, x_j)\}_{ij} = \{< \phi(x_i), \phi(x_j) >\}_{ij} = \Phi(X)\Phi(X)^T, \tag{4}$$

where $\Phi(X) = [\phi(x_1), \phi(x_2), \cdots, \phi(x_n)]^T$. Here, the kernel matrix K is the Gram matrix of the $\phi(x_i)$'s.

3.2. Principal Component Analysis and Kernel Principal Component Analysis

In standard PCA, we seek an orthogonal transformation matrix A satisfying

$$\underset{n\times d}{T} = \underset{n\times p}{X} \underset{p\times d}{A}, \tag{5}$$

where $T = [t_1, t_2, \cdots, t_d]$ for some $d \leq p$, such that each column vector t_i successively inherits maximal proportion of variance from the column vectors $x^{(l)}$'s, while ensuring the projection directions are perpendicular. The solutions can be expressed as the eigenvalue problem

$$\Sigma a_i = \lambda_i a_i, \tag{6}$$

where a_i is the i-th column of A, for $i = 1, \ldots, d$.

Following the work of [65], PCA can be extended to KPCA by first choosing a Mercer kernel k, with which x_i is transformed to $\phi(x_i)$. This maps the features in X to $\Phi(X)$. Assume that $\sum_{i=1}^{n} \phi(x_i)$ is a vector with 0 in each entry. With the Gram matrix $K = \Phi(X)\Phi(X)^T$ as defined in (4) and through the kernel trick from (3), we have the eigenvalue problem

$$K a_i^* = \lambda_i^* a_i^*, \tag{7}$$

where d is the desired dimension and $a_i^*, \cdots, a_d^*$ are the eigenvectors of K, with associated eigenvalues $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_d^*$. Hence, the advantage of the kernel-based approach is to calculate the Gram matrix K without an explicit expression for ϕ. Without the centralization assumption on ϕ, the K matrix in (7) can be replaced by

$$K^* = H_n K H_n, \tag{8}$$

where d is the desired dimension, $H_n = I_n - \frac{1}{n}1_n$, I_n is an identity matrix with dimension $n \times n$, and 1_n is a matrix of 1's with dimension $n \times n$.

We note that H_n is idempotent, since it is a square matrix satisfying $H_n = H_n H_n$. For any square matrix S with dimension $n \times n$, the average of each column of the matrix $H_n S$ is 0, as is the average of each row of the matrix SH_n. Thus, the K^* matrix is the "centralized" version of the original K matrix.

3.3. Supervised Kernel Principal Component Analysis

PCA and KPCA are unsupervised methods, since they do not consider the response variable, only considering directions of maximum variability in the covariates. If the goal is classification, this may not be ideal, since the principal components may be unrelated to the class difference. SKPCA is a supervised generalization of KPCA, which aims to find the principal components with maximal dependence on the response variable. Drawing from [34] and [35], we formulate SKPCA as follows.

In SKPCA, class information is incorporated by maximizing the Hilbert Schmidt independence criterion (HSIC) [66]. With the aforementioned reproducing kernel Hilbert spaces $\mathcal{F}$ on $\mathcal{X}$ and $\mathcal{G}$ on $\mathcal{Y}$ and related kernel functions $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$ respectively, the HSIC can be expressed as

$$\begin{aligned} HSIC(P_{\mathcal{X},\mathcal{Y}}, \mathcal{F}, \mathcal{G}) = E_{x,x',y,y'}[k(x,x')l(y,y')] + E_{x,x'}[k(x,x')]E_{y,y'}[l(y,y')] \\ - 2E_{x,y}\big(E_{x'}[k(x,x')]E_{y'}[l(y,y')]\big), \end{aligned} \tag{9}$$

where $E_{x,x',y,y'}$ represents the expectation on independent pairs of (x,y) and (x',y') (with respect to $P_{\mathcal{X},\mathcal{Y}}$) and $E_{x,x'}$ and alike are the expectations based on various marginal distributions from $P_{\mathcal{X},\mathcal{Y}}$.

With the results from [66], an empirical estimator of (9) is

$$HSIC(X, Y, \mathcal{F}, \mathcal{G}) = \frac{1}{(n-1)^2} tr(KH_n L H_n), \tag{10}$$

where K and H_n are defined as before for KPCA and $L = \{1(y_i = y_j)\}_{ij}$ is a link matrix with dimension $n \times n$, where $1(\cdot)$ is an indicator function with value 1 if the event is true and 0 otherwise.

Similarly as for KPCA, K and L can be adjusted to satisfy the centralization assumption. As discussed previously, H_n is an idempotent matrix. Therefore, following from (10),

$$\begin{aligned} HSIC^*(X, Y, \mathcal{F}, \mathcal{G}) =& \frac{1}{(n-1)^2} tr(KH_n H_n L H_n H_n) \\ =& \frac{1}{(n-1)^2} tr(H_n K H_n H_n L H_n) \\ =& \frac{1}{(n-1)^2} tr(K^* L^*), \end{aligned} \tag{11}$$

where K^* and L^* are the "centralized" versions of the K and L matrices respectively.

On another note, in the binary gender classification problem, $\text{rank}(L) = 2$ and $\text{rank}(KH_n L K H_n) \leq 2$ [35]. Therefore, we modify the link matrix according to [35] by

$$L = \{1(y_i = y_j) \times k(x_i, x_j)\}_{ij}. \tag{12}$$

It can be shown that maximization of (10) is equivalent to solving the generalized eigenvalue problem

$$Av_i = \lambda_i K v_i, \tag{13}$$

where $A = KH_nLH_nK$ and each v_i is an eigenvector with related eigenvalue λ_i for $i = 1, \cdots, d$, where d is the desired dimension [35]. Therefore, the main advantage of the link matrix in (12) becomes apparent: the rank of KH_nLKH_n may increase, permitting more eigenvalues to be computed.

3.4. Linear Discriminant Analysis and Kernel Linear Discriminant Analysis

Given a dataset with finite classes, LDA aims to find the best set of features to discriminate among the classes. We first review standard LDA, then generalize to KLDA. We note that sometimes parametric assumptions for LDA are made, such as that observations from each class are normally distributed with common covariance. Here, we make no such assumptions. Suppose that each observation x_i for $i = 1, \cdots, n$ belongs to exactly one of C classes. Define the following feature vectors: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ as the overall mean and $\bar{x_c} = \frac{1}{n_c}\sum_{i=1}^{n} x_i 1(x_i \in \text{class c})$ as the mean of the c-th class with n_c the size of the c-th class in the sample, for $c = 1, \cdots, C$.

In standard LDA, we seek to maximize the objective function

$$J(v) = \frac{v^T S_B v}{v^T S_W v}, \tag{14}$$

where v is a p x 1 vector, S_B is the between-class scatter matrix, and S_W is the within-class scatter matrix defined by

$$\begin{aligned} \underset{p\times p}{S_B} &= \sum_c n_c(\bar{x_c} - \bar{x})(\bar{x_c} - \bar{x})^T \text{ and} \\ \underset{p\times p}{S_W} &= \sum_c \sum_{i\in c} (x_i - \bar{x_c})(x_i - \bar{x_c})^T. \end{aligned} \tag{15}$$

Hence, maximizing $J(v)$ involves finding some rotation of the scatter matrices such that the "distance" between the groups is maximized relative to the variations within each group.

Maximization of $J(v)$ in (14) is equivalent to solving the generalized eigenvalue problem

$$S_B v_i = \lambda_i S_W v_i, \tag{16}$$

where each v_i is an eigenvector with corresponding eigenvalue λ_i, for $i = 1, \cdots, d$, where d is the desired dimension.

LDA is generalized to KLDA using the kernel representation from (3). Analogously to LDA above, we seek a solution v^* that will result in the maximization of the objective function

$$J(v) = \frac{v^T S_B^* v}{v^T S_w^* v}, \tag{17}$$

where now $v \in \mathcal{F}$ and S_B^* and S_W^* are the between-class and within-class scatter matrices in $\mathcal{F}$ defined by

$$\begin{aligned} m^\phi &= \frac{1}{n}\sum_{i=1}^{n} \phi(x_i), \\ m_c^\phi &= \frac{1}{n_c}\sum_{i=1}^{n} \phi(x_i) 1(x_i \in \text{class c}), \\ S_B^* &= \sum_c n_c(m_c^\phi - m^\phi)(m_c^\phi - m^\phi)^T, \text{ and} \\ S_W^* &= \sum_c \sum_{i\in c} (\phi(x_i) - m_c^\phi)(\phi(x_i) - m_c^\phi)^T. \end{aligned} \tag{18}$$

The above expressions involve knowledge of ϕ, which is often not available. It can be shown that equation (17) is equivalent to

$$J(u) = \frac{u^T M u}{u^T N u}, \tag{19}$$

where

$$\begin{aligned} M_c = (M_{cj})_j &= \left(\frac{1}{n_c} \sum_{h=1}^{n} k(x_j, x_h) 1(x_h \in \text{class c}) \right)_j, \\ \bar{M} = (\bar{M}_j)_j &= \left(\frac{1}{n} \sum_{h=1}^{n} k(x_j, x_h) \right)_j, \\ M &= \sum_c n_c (M_c - \bar{M})(M_c - \bar{M})^T, \\ K_c &= K \times outer(X, X_c), \\ N &= \sum_c K_c H_{n_c} K_c^T, \end{aligned} \tag{20}$$

X_c is a matrix of dimension $n_c \times p$ with rows being features from the c-th class, and $outer(X, X_c)$ is an $n \times n_c$ matrix with its ij-th element as $1(x_i$ is the j-th observation in class c$)$. A full discussion of KLDA can be found in [17].

Maximization of $J(u)$ in equation (19) is equivalent to solving the generalized eigenvalue problem

$$Mu_i = \lambda_i N u_i, \tag{21}$$

where each u_i is an eigenvector with associated eigenvalue λ_i, for $i = 1, \cdots, d$ with d as the desired dimension.

Comparing the generalized eigenvalue problems in (16) and (21), the structures of matrices S_B and M are similar, since both "measure" the variation between different classes.

Let $W_c = [w_{c,1}, \cdots, w_{c,n_c}] = K_c H_{n_c}$, a matrix of dimension $n \times n_c$. Due to the centralization function of H_{n_c}, W_c has row-sum equal to zero for every row. Besides, $K_c H_{n_c} (K_c H_{n_c})^T = W_c W_c^T = \sum_{i=1}^{n_c} w_{c,i} w_{c,i}^T$. For the matrix N, due to the idempotent property of H_{n_c},

$$N = \sum_c K_c H_{n_c} H_{n_c} K_c^T = \sum_c \sum_{i=1}^{n_c} w_{c,i} w_{c,i}^T. \tag{22}$$

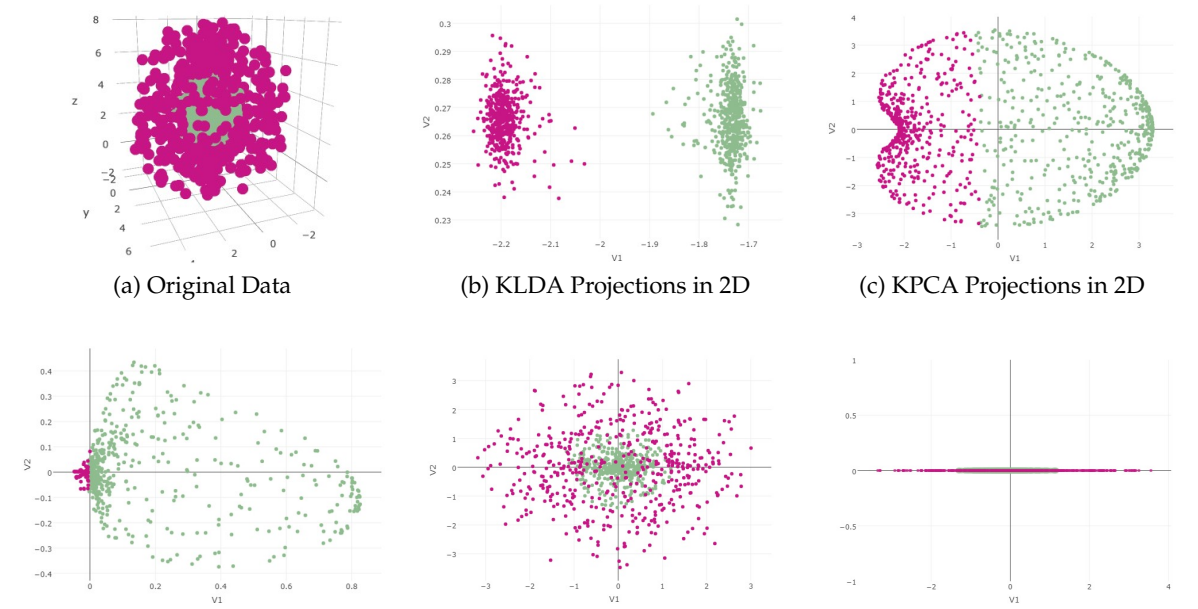
Thus, the matrix N has an identical structure to the S_W and S_W^* matrices, which "measure" the overall variation within groups.

4. Visualization on Simulation Studies

To visualize and improve understanding of the manifold learning methods KPCA, SKPCA, and KLDA, we apply them in three simulation studies. For comparison, the linear techniques PCA and LDA are also considered. Each dataset contains nonlinear patterns, and the goal is to transform the data to be linearly separable. For this reason, the radial basis function (RBF)

$$k(x_i, x_j) = e^{-\delta ||x_i - x_j||_2^2} \tag{23}$$

is chosen as a kernel for each pair of observed vectors x_i, x_j. For each DR method, a range of values for the tuning parameter δ are tested and selected to visually separate the classes. A full discussion of the RBF kernel, among others, can be found in [67]. Figures 1, 2, and 3 compare the original data to 2-dimensional projections from each DR method. In each plot, color corresponds to the true class to which an observation belongs.

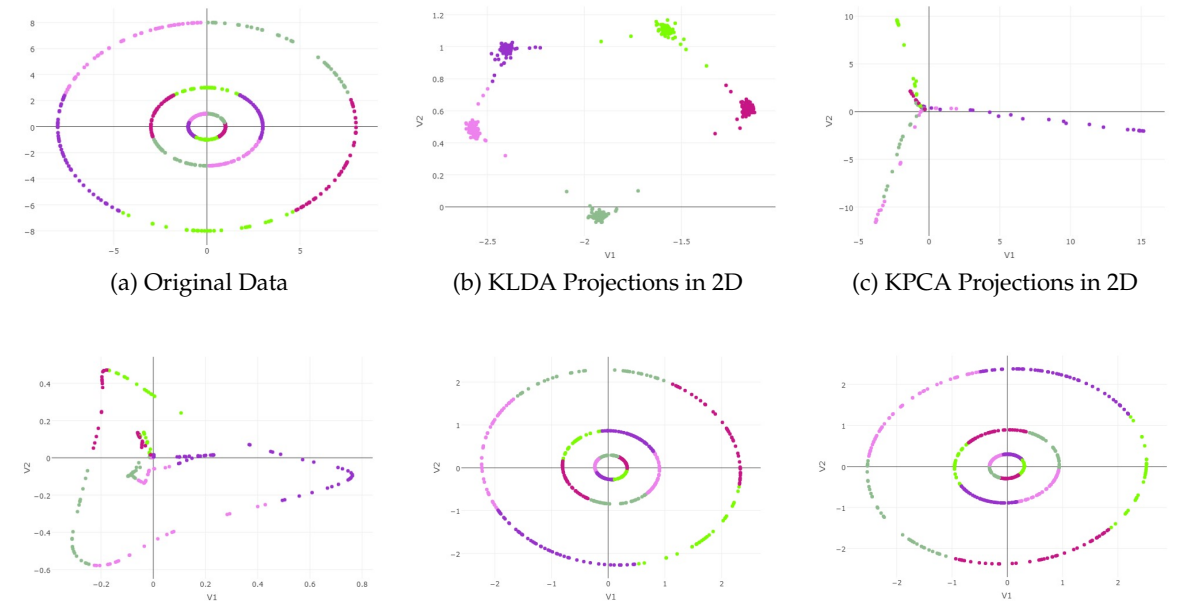


(a) Original Data (b) KLDA Projections in 2D (c) KPCA Projections in 2D

(d) SKPCA Projections in 2D (e) PCA Projections in 2D (f) LDA Projections in 2D

Figure 1. Wine Chocolate Simulation Study.

In the first simulation study, the original data are plotted in 3D in Figure 1(a); the green sphere is embedded within the magenta group, necessitating nonlinear manifold learning. The KLDA projections in (b) are linearly separable with very good variation between the classes and a fair amount of variation within the classes. KPCA and SKPCA projections in (c) and (d) are at least approximately linearly separable, as it is not clear whether there is a linear boundary that perfectly separates the two classes. In (e), PCA fails to linearly separate the groups, rotating the wine chocolate in 2D. The maximum dimension LDA can retain is $p - 1$; with 2 classes, the projections must be plotted on a 1D number line, given in (f). Points from the two classes overlap considerably in plots (e) and (f).

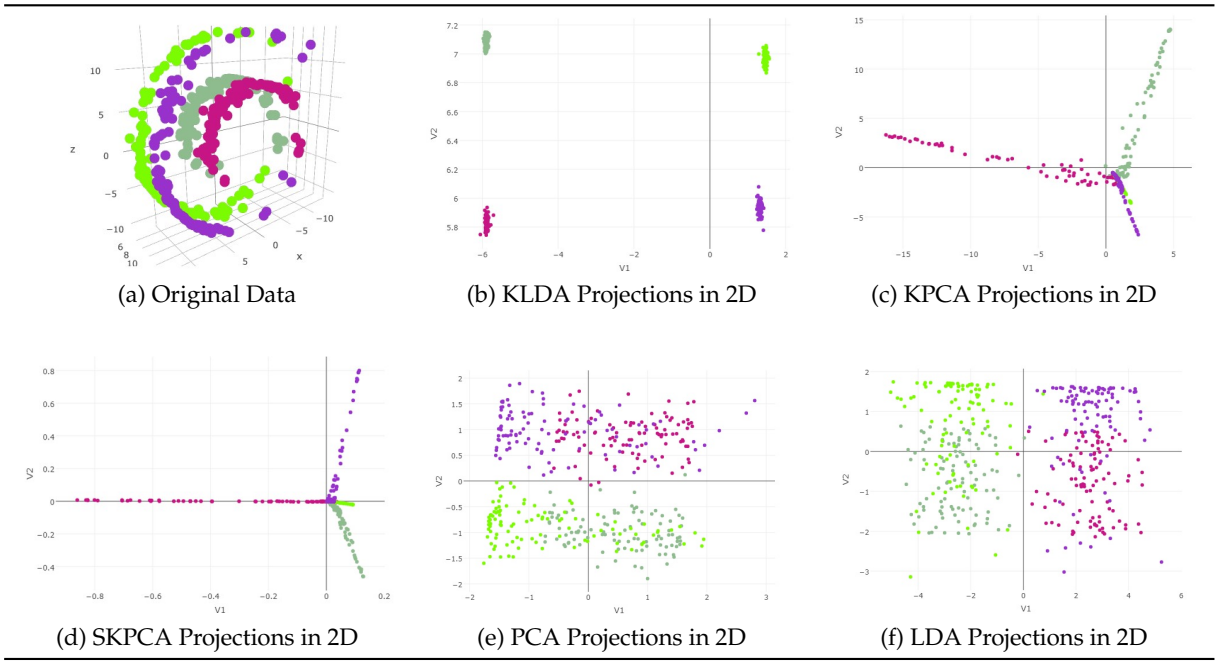


(a) Original Data (b) KLDA Projections in 2D (c) KPCA Projections in 2D

(d) SKPCA Projections in 2D (e) PCA Projections in 2D (f) LDA Projections in 2D

Figure 2. Apple Tart Simulation Study.

In the second simulation study, the original data in Figure 2(a) follow a nonlinear pattern. In (b), KLDA produces groups which are linearly separable. The KPCA projections are approximately linearly separable in (c); however, there is some overlap between groups, especially the green and pink groups in the third quadrant. In (d), SKPCA produces almost linearly separable groups. In plots (e) and (f), PCA and LDA simply rotate the original data in 2D space, as expected.



(a) Original Data (b) KLDA Projections in 2D (c) KPCA Projections in 2D

(d) SKPCA Projections in 2D (e) PCA Projections in 2D (f) LDA Projections in 2D

Figure 3. Swiss Roll Simulation Study.

For the third simulation study, the original data in Figure 3(a) are in 3D and follow a swirling, nonlinear pattern. In (b), KLDA yields favorable results; the groups are well-separated linearly. KPCA and SKPCA in (c) and (d) also produce good results, although in (c) more separation between the purple and bright green groups would be ideal. In (e) and (f), respectively, PCA and LDA merely rotate the original data projected in 2D space; there is no linear separation between the magenta and purple groups, nor between the two green groups.

For all three simulation studies, KLDA, KPCA, and SKPCA are effective to transform the data into linearly separable groups. In all cases, the projected data become approximately linearly separable after applying KLDA, KPCA, or SKPCA. In general, KLDA and SKPCA perform the best here. Their success over KPCA is expected, since KLDA and SKPCA are supervised techniques. On the other hand, results indicate that KPCA and SKPCA are more sensitive than KLDA to different choices of tuning parameters. Hence, SKPCA and KPCA may perform better for alternative choices of parameters. In all our studies, the nonlinear techniques outperform linear PCA and LDA. These preliminary studies suggest the radial kernel is appropriate for our face analysis experiments.

5. Kernel-based Dimension Reduction Optimization and Classification on Morph-II

Motivated by the nonlinear nature of facial demographic analysis, we propose and implement a novel machine learning process for the Morph-II dataset. We consider the kernel-based DR methods KPCA, SKPCA, and KLDA on three types of appearance-based extracted features (LBP, BIF, and HOG) for the gender classification task. We illustrate parameter optimization and compare the performance of these methods on Morph-II.

5.1. Longitudinal Face Database

MORPH [68] is one of the most popular face databases available to the public, especially for age estimation, race classification, and gender classification. Multiple versions of MORPH have been released, and the version adopted in this work is the 2008 MORPH-II non-commercial release (referred to as Morph-II in this paper). Morph-II includes over 55,000 mugshots with longitudinal spans and metadata such as date of birth, race, gender, and age.

In addition to its size, Morph-II presents challenges because of disproportionate race and gender ratios. About 84.6% of images are of males, while only about 15.4% of images are of females. Imbalanced classes are known to negatively affect certain classification algorithms [69]. Moreover, Morph-II is skewed in terms of race, with approximately 77.2% of images picturing black subjects. Guo et al. found that age, gender, and race interact for demographic analysis tasks including gender classification, race classification, and age prediction [48,60,70], so both race and gender imbalance in Morph-II can hamper gender classification.

5.2. Subsetting Scheme

To overcome the uneven race and gender distributions in Morph-II, Guo et al. proposed a subsetting scheme [48]. Since then, many studies on Morph-II have adopted such an evaluation protocol. Based on discussions in Guo et al. [48], a new automatic subsetting scheme is proposed in [71], aiming to automatically ensure independent training and testing sets. Additionally, inconsistencies in age, gender, and race in Morph-II have been identified and corrected in [71]. After following the steps to clean MORPH-II outlined in [71], we apply the automatic subsetting scheme, summarized in Figure 4 and described below.

Let W be the Whole Morph-II dataset, S the selected training/testing set, and R the remaining set. We further divide S into even subsets S_1 and S_2. Separately within each subset S_1 and S_2, we fix the ratios of white (W) to black (B) images as 1:1 and male (M) to female (F) images as 3:1. Further, S_1 and S_2 have been selected such that the age distributions within each set are similar (details shown in [71]). The gender and race summaries for the subsetting scheme are shown in Table 2. Most importantly, the sets R, S_1, and S_2 are independent; no sets share images from the same subject. We use S as an alternating training and testing set. First, we train on S_1 and test on $S_2 \cup R$, then we train on S_2 and test on $S_1 \cup R$. The final classification accuracy is obtained by averaging the classification accuracies from the alternations.

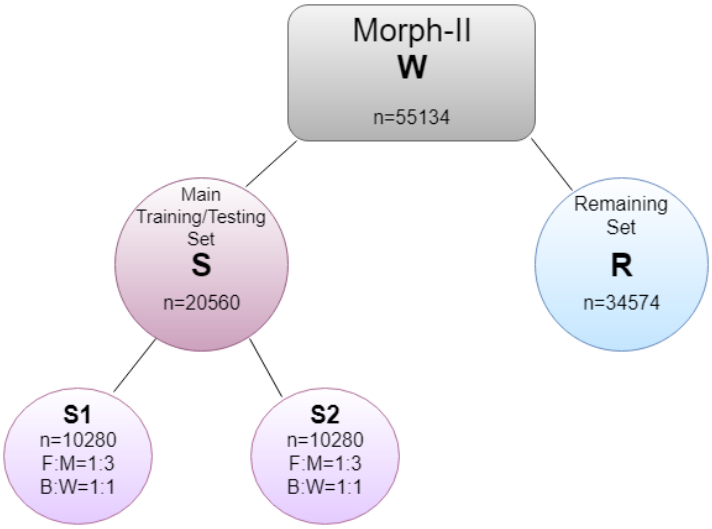


Figure 4. Flowchart representing our subsetting scheme [71] for MORPH-II, which improves the one from [48].

Table 2. Number of Images in Subsets by Race and Gender

	WF	BF	WM	BM	dF	dM	Overall	F	M
S1	1,285	1,285	3855	3,855	0	0	10,280	2570	7,710
S2	1,285	1,285	3,855	3,855	0	0	10,280	2,570	7,710
R	0	3,150	220	28,980	144	1,850	34,344	3,294	31,050
Overall	2,570	5,720	7,930	36,690	144	1,850	54,904	8,434	46,470

Race-gender combinations are abbreviated, e.g., **BF** represents the black female group. Abbreviations **dF** and **dM** represent those who are neither black nor white in race.

5.3. Facial Feature Extraction

In computer vision, image preprocessing is often an essential first step to reduce unnecessary variation, decrease pixel dimension, and simplify pixel encoding. Despite the standard format of police photography in mugshots, Morph-II photographs vary in head-tilt, camera distance, occlusion, and illumination. We address this variation as follows. Images are first converted to grayscale. Next, faces are automatically detected, eliminating the background and hair, so that no external cues can be used to classify gender. The resulting images are centered and scaled with respect to the center of the irises. Finally, the images are cropped to be 70 pixels tall by 60 pixels wide. Full methodological details are given in [72] and align with standard preprocessing protocols from face analysis.

After preprocessing, pixel-related features are extracted from the face images in Morph-II. As discussed previously, there are numerous approaches for feature extraction. In this study on Morph-II, we incorporate domain expertise by choosing three well-established appearance-based models from image analysis: local texture techniques such as local binary patterns (LBP) [55–58], biologically-inspired features (BIF) [60,61], and histogram of oriented gradients (HOG) [60]. Additionally, these model-based approaches provide "robust interpretation ... by constraining solutions to be face-like" [73]. Detailed documentation of our feature extraction process can be found in [72,74].

Table 3. Parameter Summary

Features	LBP	$s = 10, 12, 14, 16, 18, 20$ $r = 1, 2, 3$
	HOG	$s = 4, 6, 8, 10, 12, 14$ $o = 4, 6, 8$
	BIF	s=$7 - 37, 15 - 29$ $\gamma = 0.1, 0.2, \ldots, 1.0$
Dimension Reduction	KPCA	$\delta = \pm 0.1, \pm 1, \pm 5, \pm 10, \pm 100$
	SKPCA	δ =-0.0001,-0.001 $\eta = 0.001, 0.01, 0.1, 1$
	KLDA	$\delta = \pm 0.01, \pm 0.1, \pm 1, \pm 5, \pm 10, \pm 100$
Classifier	Linear SVM	$c = 10^{-8}, \ldots, 10^{-1}, 1, 10, \ldots, 10^{8}$

5.4. Kernel-Based Dimension Reduction Optimization

Tuning parameter selection is essential for kernel-based methods in order to achieve good results. Within the framework of feature extraction, dimension reduction, and the classification model, there are many combinations of parameters to be considered. The main parameters and tested values are summarized in Table 3 and discussed as follows. LBP features have two main parameters: block size s and window radius r. For HOG, the two main parameters are block size s and number of orientations o. For BIF, we consider the block size s and the parameter γ, which represents the spatial aspect ratio; there is also a choice of pooling operation, which we select here as the standard deviation operation.

For each dimension reduction method, the radial kernel

$$k(x_i, x_j) = e^{\delta ||x_i - x_j||_2^2} \tag{24}$$

is used for each pair of observation vectors x_i, x_j, based on the results from our simulation studies. In the kernel, we must select the tuning parameter δ, which scales the extent of similarity between pairs of vectors. This parameter must be chosen with particular care, since a poor choice can result in transformed features with little to no variability. Empirically, we observed that SKPCA was more sensitive than KLDA and KPCA to the choice of δ; values of δ at or above 1 resulted in a rank deficient matrix and failure to compute all requested eigenvalues. For SKPCA, we consider an additional scaling parameter η in the modified link function proposed by Wang et al. [35]:

$$l(y_i, y_j) = e^{\eta\delta||x_i - x_j||_2^2}, \tag{25}$$

for all observed responses y_i, y_j in the same class. The scale parameter η enables the weighing of dependence between the covariates and response.

Finally, we choose a linear SVM to classify gender based on the dimension-reduced, transformed features. The motivation for this classifier is discussed in the next section. The main parameter for linear SVM is the cost c, which measures the extent to which misclassification in training will be permitted. We consider values of c from 10^{-8} to 10^{8}.

Table 4. Tuning Results on a Subset of MORPH-II

Method	Feature	Parameters	Accuracy
KPCA	BIF $s = 7-37, \gamma = 0.1$	$\delta = -1, c = 10$	0.882
	BIF $s = 7-37, \gamma = 0.6$	$\delta = -1, c = 10$	0.882
	BIF $s = 15-29, \gamma = 0.1$	$\delta = -1, c = 100$	0.882
	BIF $s = 15-29, \gamma = 0.6$	$\delta = -1, c = 10$	0.882
	HOG $s = 4, o = 4$	$\delta = -100, c = 0.1$	0.917
	HOG $s = 4, o = 4$	$\delta = -5, c = 0.001$	0.919
	HOG $s = 4, o = 4$	$\delta = -1, c = 0.001$	0.917
	HOG $s = 4, o = 4$	$\delta = -0.1, c = 0.1$	0.917
	LBP $s = 10, r = 1$	$\delta = -100, c = 0.1$	0.912
	LBP $s = 10, r = 1$	$\delta = -5, c = 0.1$	0.912
	LBP $s = 10, r = 1$	$\delta = -1, c = 0.001$	0.912
	LBP $s = 10, r = 1$	$\delta = -0.1, c = 0.1$	0.912
SKPCA	BIF $s = 7-37, \gamma = 0.2$	$\delta = 0.0001, \eta = 0.1, c = 1$	0.899
	BIF $s = 7-37, \gamma = 0.8$	$\delta = 0.0001, \eta = 0.1, c = 1$	0.899
	BIF $s = 15-29, \gamma = 0.5$	$\delta = 0.0001, \eta = 0.1, c = 1$	0.899
	HOG $s = 6, o = 6$	$\delta = 0.0001, \eta = 0.001, c = 1$	0.931
	HOG $s = 6, o = 6$	$\delta = 0.0001, \eta = 0.01, c = 0.001$	0.931
	HOG $s = 6, o = 6$	$\delta = 0.0001, \eta = 0.1, c = 0.001$	0.931
	LBP $s = 14, r = 2$	$\delta = 0.0001, \eta = 0.001, c = 1$	0.937
	LBP $s = 14, r = 2$	$\delta = 0.0001, \eta = 0.01, c = 1$	0.937
	LBP $s = 14, r = 2$	$\delta = 0.0001, \eta = 0.1, c = 1$	0.938
	LBP $s = 14, r = 2$	$\delta = 0.0001, \eta = 1, c = 1$	0.939
KLDA	BIF $s = 7-37, \gamma = 0.3$	$\delta = -1, c = 10$	0.875
	BIF $s = 7-37, \gamma = 0.6$	$\delta = -1, c = 100$	0.875
	BIF $s = 15-29, \gamma = 0.2$	$\delta = -1, c = 10$	0.875
	BIF $s = 15-29, \gamma = 0.8$	$\delta = -1, c = 100$	0.875
	HOG $s = 4, o = 4$	$\delta = 1, c = 1$	0.917
	HOG $s = 4, o = 6$	$\delta = 1, c = 1$	0.917
	HOG $s = 12, o = 8$	$\delta = -1, c = 100$	0.904
	LBP $s = 10, r = 1$	$\delta = -0.1, c = 1$	0.906
	LBP $s = 10, r = 1$	$\delta = 1, c = 1$	0.908
	LBP $s = 14, r = 1$	$\delta = 0.1, c = 10$	0.898

We tune on small subsets of Morph-II to reduce runtime, memory consumption, and risk of over-fitting. 1000 images from S_1 and 1000 images from S_2 are randomly selected. The standard method of grid search is followed for tuning on these subsets. For each combination of parameters, a model is trained on the subset from S_1 and then tested on the subset from S_2. For each dimension

reduction method paired with each feature type (BIF, HOG, and LBP), the best three or four accuracy rates from tuning are obtained. (Except in the case of ties, we choose only the top three accuracy rates.) The tuning results for these top-performing parameters are given in Table 4. The parameters corresponding to these maximum accuracy rates are applied to the full dataset through the previously discussed evaluation protocol. Although this protocol involves testing on images from S_1 and S_2, any overlap of images is minor (in each testing set, less than 2.3% of images have been used in tuning) and has little impact on the reported accuracy (discussed in Section 6). Regardless, the tuning parameters are selected through the same procedure, so the classification performances can be fairly compared among all considered DR methods.

Table 5. Gender Classification Results on MORPH-II

Method	Feature	Parameters	Acc(1)	TPR(2)	TNR(3)	Mem(4)	Time(5)
KPCA	BIF $s=7-37,\gamma=0.1$	$\delta=-1,c=10$	0.9296	0.9473	0.8127	34.04	42.26
	BIF $s=7-37,\gamma=0.6$	$\delta=-1,c=10$	**0.9297**	0.9455	0.8112	34.68	36.94
	BIF $s=15-29,\gamma=0.1$	$\delta=-1,c=100$	0.9071	0.9377	0.7050	31.74	33.83
	BIF $s=15-29,\gamma=0.6$	$\delta=-1,c=10$	0.9096	0.9374	0.7266	31.80	35.97
	HOG $s=4,o=4$	$\delta=-100,c=0.1$	**0.9391**	0.9726	0.7172	34.00	31.54
	HOG $s=4,o=4$	$\delta=-5,c=0.001$	**0.9391**	0.9727	0.7170	34.00	30.86
	HOG $s=4,o=4$	$\delta=-1,c=0.001$	**0.9391**	0.9724	0.7192	34.00	32.17
	HOG $s=4,o=4$	$\delta=-0.1,c=0.1$	0.9364	0.9626	0.7634	34.35	31.41
	LBP $s=10,r=1$	$\delta=-100,c=0.1$	**0.9391**	0.9726	0.7172	34.00	31.54
	LBP $s=10,r=1$	$\delta=-5,c=0.1$	**0.9391**	0.9726	0.7172	34.00	30.86
	LBP $s=10,r=1$	$\delta=-1,c=0.001$	**0.9391**	0.9724	0.7192	34.00	32.17
	LBP $s=10,r=1$	$\delta=-0.1,c=0.1$	0.9364	0.9626	0.7634	34.35	31.41
SKPCA	BIF $s=7-37,\gamma=0.2$	$\delta=0.0001,\eta=0.1,c=1$	0.9507	0.9616	0.8781	35.48	42.04
	BIF $s=7-37,\gamma=0.8$	$\delta=0.0001,\eta=0.1,c=1$	**0.9532**	0.9639	0.8823	33.04	38.34
	BIF $s=15-29,\gamma=0.5$	$\delta=0.0001,\eta=0.1,c=1$	0.9260	0.9477	0.7827	20.03	34.58
	HOG $s=6,o=6$	$\delta=0.0001,\eta=0.001,c=1$	0.9467	0.9645	0.8292	36.69	37.39
	HOG $s=6,o=6$	$\delta=0.0001,\eta=0.01,c=0.001$	**0.9489**	**0.9786**	0.7528	38.28	53.96
	HOG $s=6,o=6$	$\delta=0.0001,\eta=0.1,c=0.001$	0.9488	**0.9786**	0.7517	39.83	60.55
	LBP $s=14,r=2$	$\delta=0.0001,\eta=0.001,c=1$	**0.9585**	0.9727	0.8641	28.68	25.33
	LBP $s=14,r=2$	$\delta=0.0001,\eta=0.01,c=1$	**0.9585**	0.9764	0.8642	23.22	38.42
	LBP $s=14,r=2$	$\delta=0.0001,\eta=.1,c=1$	**0.9585**	0.9730	0.8640	29.87	28.00
	LBP $s=14,r=2$	$\delta=0.0001,\eta=1,c=1$	**0.9585**	0.9727	0.8640	27.92	22.83
KLDA	BIF $s=7-37,\gamma=0.3$	$\delta=-1,c=10$	0.9415	0.9539	0.8594	24.89	34.50
	BIF $s=7-37,\gamma=0.6$	$\delta=-1,c=100$	**0.9426**	0.9558	**0.8858**	24.74	35.46
	BIF $s=15-29,\gamma=0.2$	$\delta=-1,c=10$	0.9131	0.9374	0.7532	22.80	26.78
	BIF $s=15-29,\gamma=0.8$	$\delta=-1,c=100$	0.9205	0.9421	0.7783	22.83	33.88
	HOG $s=4,o=4$	$\delta=1,c=1$	0.9369	0.9517	0.8392	36.52	81.71
	HOG $s=4,o=6$	$\delta=1,c=1$	**0.9398**	0.9545	0.8425	52.48	148.24
	HOG $s=12,o=8$	$\delta=-1,c=100$	0.9175	0.9421	0.7542	21.18	21.57
	LBP $s=10,r=1$	$\delta=-0.1,c=1$	**0.9418**	0.9578	0.8428	24.58	37.17
	LBP $s=10,r=1$	$\delta=1,c=1$	0.9417	0.9558	0.8486	24.70	36.45
	LBP $s=14,r=1$	$\delta=0.1,c=10$	0.9392	0.9543	0.8397	20.77	31.12

(1) Acc represents mean accuracy.
(2) TPR represents mean true positive rate (recall/sensitivity): the proportion of male faces correctly classified.
(3) TNR represents mean true negative rate (specificity): the proportion of female faces correctly classified.
(4) Mem represents mean memory in gigabytes.
(5) Time represents mean runtime in hours for training and testing.

5.5. Gender Classification

For the classification part of the modeling, linear SVM is adopted. Many face analysis studies have involved SVM, as summarized in [75]. Briefly, SVM identifies a separating hyperplane with maximal margin between the classes. Several popular kernels for SVM include linear, polynomial, and RBF [67]. We select the linear kernel, because directions of variability in the data are expected to be linear after the nonlinear transformations of KPCA, SKPCA, or KLDA. Indeed, Schölkopf et al. observed this to be true for KPCA in their landmark study [65]. The linear kernel for SVM also reduces computational cost, compared to nonlinear kernels.

With the parameters in Table 4 that are selected from tuning on subsets, we implement dimension reduction and classification on the full Morph-II dataset, following the subsetting scheme discussed in Section 5.2. The challenges of the large size of Morph-II, the high dimensionality of the features, and the computational complexity of these dimension reduction methods necessitate the use of high-performance computing (HPC). For example, the kernel matrix for each dimension reduction method is 55134×55134, requiring approximately 23 gigabytes of storage. Thus, we implement the process on the HiPerGator 2.0 supercomputing cluster at the University of Florida. The code is written in R. The R package *rARPACK* is used to optimize the solving of eigenvalue problems [76], and the *e1071* package is utilized for training and testing the SVM model [77].

6. Experiment Results

The kernel-based DR methods KPCA, SKPCA, and KLDA are applied to three facial feature extraction methods: BIF, HOG, and LBP. The DR methods transform the feature data, then reduce the dimension. In all cases, a dimension of 100 is retained, substantially lower than the dimension of the original feature space. The dimensionality of 100 is selected as a trade-off between computation time and classification accuracy based on our preliminary studies. The transformed and dimension-reduced data serve as input for the linear SVM, which classifies each image subject as male or female. Additionally, these predicted gender classes are mapped to probabilities through a sigmoid function, following [78]. This process is applied to each alternation of the evaluation protocol: 1) train on S_1, test on $S_2 \cup R$ and 2) train on S_2, test on $S_1 \cup R$. The classification results are averaged over these two testing sets. The mean classification accuracy over the testing images is chosen as the evaluation criterion for our methods on Morph-II, as it is the usual performance metric for gender classification [60], especially in similar studies [49,51,52,79].

These mean classification results from Morph-II are shown in Table 5. In addition to the accuracy, the true positive rate (also known as sensitivity or recall) and true negative rate (also called specificity) are given. For this study, we define the true positive rate (TPR) as the proportion of male faces correctly classified, while the true negative rate (TNR) as the proportion of female faces correctly classified. The memory and runtime are also listed in Table 5. The runtime is the total time for training and testing on HPC, i.e., the average of time1 (train on S_1, test on $S_2 \cup R$) and time2 (train on S_2, test on $S_1 \cup R$). As mentioned in Section 5.4, there is a small overlap between the tuning and testing sets that could contribute to over-fitting. We have assessed the potential impact of over-fitting on our reported accuracy rates and found it to be very small: it is estimated to be (at most) between 0.09% and 0.2% and to monotonically decrease as reported accuracy rates increase.

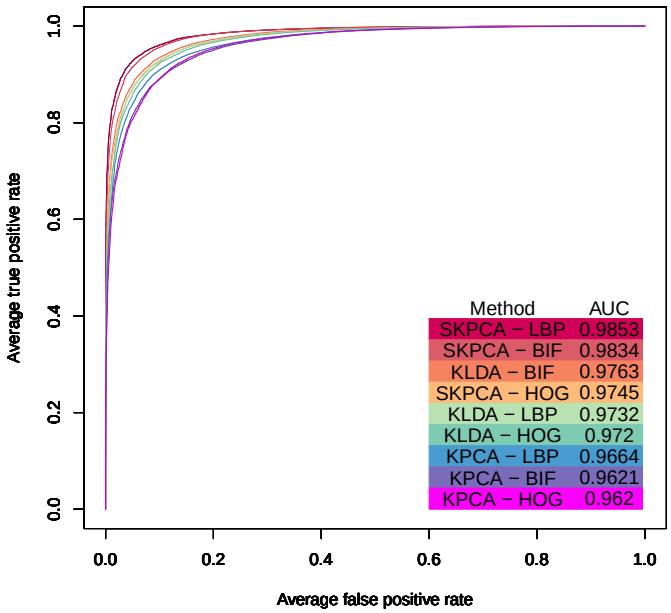


Figure 5. Receiver operating characteristic (ROC) curve and area under the curve (AUC) are compared by method for gender classification on Morph-II. Each color corresponds to a DR method paired with feature type. For each probability threshold, the true and false positive rates are reported as the averages from the testing sets of the alternating evaluation protocol.

The classification performance is further visualized in Figure 5 through receiver operating characteristic (ROC) curves for the nine combinations of DR method and feature extraction type. For each combination, its displayed curve corresponds to the "best" results from Table 5 (the combination of parameters reaching maximum mean classification accuracy or maximum mean true positive rate in the event of ties). For each alternation of the evaluation protocol, the true and false positive rates in testing are calculated for each probability threshold. To construct the ROC curves, each of the resulting rates for each threshold is averaged over the testing sets.

Table 5 shows that for the feature BIF, SKPCA and KLDA outperform KPCA. For the feature HOG, SKPCA achieves higher accuracy than both KPCA and KLDA, while the latter two techniques perform very similarly. Last, for the feature LBP, SKPCA produces better classification accuracy than KPCA and KLDA. In summary, our experiment's results indicate that SKPCA outperforms KLDA consistently, while KLDA outperforms KPCA for all three features BIF, LBP, and HOG. On the other hand, for KPCA, the features HOG and LBP produce approximately the same accuracies, outperforming BIF. For SKPCA, LBP achieves slightly better results than BIF, while LBP and BIF both outperform HOG. Finally, for KLDA, BIF reaches slightly higher accuracy than LBP, while BIF and LBP both exceed HOG.

In most cases, the accuracy (in Table 5) and AUC (in Figure 5) metrics agree on the best methods. An exception is that SKPCA with the HOG features achieves slightly higher accuracy (94.89%) than KLDA with the BIF features (94.18%), but SKPCA with HOG has lower AUC than KLDA with BIF. The other exception is that KPCA with the HOG features has the lowest AUC of the nine methods, but its accuracy is higher than KPCA with the BIF features. In summary, the accuracy and AUC results imply that SKPCA generally performs best for gender classification on Morph-II, while KLDA tends to outperform KPCA. Meanwhile, the LBP and BIF features often yield better classification performance, with less memory usage, than the HOG features.

It is interesting that, overall, LBP achieves even slightly better performance than BIF for the dimension reduction method SKPCA on the task of gender classification, since BIF is popular in demographic analysis such as age estimation, gender classification, and race classification [48,49,60, 70,79]. Another interesting fact is displayed by the results of the true positive and negative rates in Table 5: males have a higher probability of correct identification than females, with the biggest margin

exceeding 20%. Our finding is consistent with [61]: females are more challenging to correctly classify than males, both for automatic approaches and human perception. Similarly, for race classification on Morph-II, Guo and Mu found in [70] that training a model on female faces (and testing on male faces) contributed to significantly more errors on average than training on male faces (and testing on female faces), even when controlling for differences in the training sample sizes. Our results also indicate that, overall, HOG and LBP outperform BIF for males, while BIF works consistently better than LBP and HOG for females.

Next, in Table 6 we compare our results to studies using similar methods on Morph-II, as well as recent state-of-the-art works with deep learning on MORPH-II. With the exception of [61], all studies' results in the table are mean testing classification accuracy from an alternating evaluation protocol based on Guo et al [48]. Hence, our results can be directly compared to these studies. With LBP features, SKPCA, and a linear support vector machine (SVM), our gender classification accuracies approximate 96%, competitive with benchmark results. Interestingly, several reported accuracy rates from human observers of gender range from 96% [42] to 96.9% [61]. The similarity in recognition rates between our methods and human observers can further validate the success of our approach.

Table 6. Comparison Results for Gender Classification on MORPH-II

Method	Accuracy	Reference	Year
BIF+OLPP	98%	[49]	2011
BIF+PLS	97.34%	[49]	2011
BIF+KPLS	98.2%	[49]	2011
BIF+CCA	95.2%	[79]	2014
BIF+KCCA	98.4%	[79]	2014
BIF+rCCA	97.6%	[79]	2014
Multi-scale CNN	97.9%	[52]	2014
Ranking CNN	97.9%	[51]	2015
BIF+Hierarchical-SVM	97.6%	[61]	2015
Human Estimators	96.9%	[61]	2015
LBP+SKPCA+L-SVM	95.85%	This work	2019

7. Kernel-based Dimension Reduction Optimization and Classification on FG-NET

For further comparison between KPCA, SKPCA, and KLDA, we apply a modification of our approach from Section 5 to a smaller face dataset, the face and gesture recognition network (FG-NET). FG-NET is a popular, publicly available database used for age estimation, gender classification, face recognition, and other demographic analysis tasks [80]. It contains 1002 images from 82 subjects: 47 males and 35 females with ages varying from 0 to 69 years [80].

For each image, 109 features are extracted using the Active Appearance Model (AAM), a commonly adopted appearance-based approach that models the shape and texture of the face [73,81]. As in Section 5.4, the radial kernel defined in equation (24) is chosen for each of the DR methods KPCA, SKPCA, and KLDA. Additionally, the modified link function from equation (25) is applied in the SKPCA algorithm. Thus, the tuning parameter δ in the radial kernel and η in the modified link function must be selected. As in our experiments on Morph-II, linear SVM is chosen as the classifier for FG-NET. On Morph-II, values of the cost parameter c ranging from 10^{-8} to 10^{8} were tested. On FG-NET, we have observed convergence issues in the SVM algorithm for values of c exceeding 10, so only the values $10^{-8}, 10^{-7}, \ldots, 10^{-1}, 1, 10$ are tested. The considered tuning parameters are summarized in Table 7.

For cross-validation, we use leave-one-person-out (LOPO), the most well-accepted scheme for FG-NET [80]. LOPO is a variation of k-fold cross-validation that produces independent training and testing folds in longitudinal datasets. The number of folds k is set equal to the number of subjects in the dataset, so $k = 82$ here. For $i = 1, 2, \ldots 82$, testing fold i contains only images of person i, while training fold i contains all remaining images. Similarly to on Morph-II, we choose the mean classification accuracy over the testing folds to be the evaluation criterion.

Table 7. Parameter Summary for FG-NET

Dimension Reduction	KPCA	$\delta = 3.2, 3.2\overline{6}, 3.\overline{3}, 3.4, 3.4\overline{6}, 3.5\overline{3}, 3.6, 3.\overline{6}, 3.7\overline{3}, 3.8$
	SKPCA	$\delta = 0.0098$ $\eta = 0.001, 0.01, 0.1, 1$
	KLDA	$\delta = 3, 3.\overline{5}, 4.\overline{1}, 4.\overline{6}, 5.\overline{2}, 5.\overline{7}, 6.\overline{3}, 6.\overline{8}, 7.\overline{4}, 8$
Classifier	Linear SVM	$c = 10^{-8}, \ldots, 10^{-1}, 1, 10$

Table 8. Gender Classification Results on FG-NET

Method	Parameters	Acc(1)	TPR(2)	TNR(3)
KPCA	$\delta = 3.2\overline{6}, c = 10$	**0.7025**	0.7325	0.6621
	$\delta = 3.\overline{3}, c = 10$	0.6932	0.7233	0.6528
	$\delta = 3.4, c = 10$	0.6801	0.6651	0.7001
SKPCA	$\delta = 0.0098, \eta = 0.1, c = 10$	**0.7154**	0.7542	0.6633
	$\delta = 0.0098, \eta = 1, c = 0.1$	0.6933	0.7413	0.6289
	$\delta = 0.0098, \eta = 0.01, c = 0.1$	0.6893	0.7701	0.5809
KLDA	$\delta = 3, c = 0.01$	**0.7225**	0.7593	0.6730
	$\delta = 5.\overline{7}, c = 1$	0.7176	0.7810	0.6324
	$\delta = 8, c = 0.1$	0.7131	0.7431	0.6727

(1) Acc represents mean accuracy.
(2) TPR represents mean true positive rate (recall/sensitivity): the proportion of male faces correctly classified.
(3) TNR represents mean true negative rate (specificity): the proportion of female faces correctly classified.

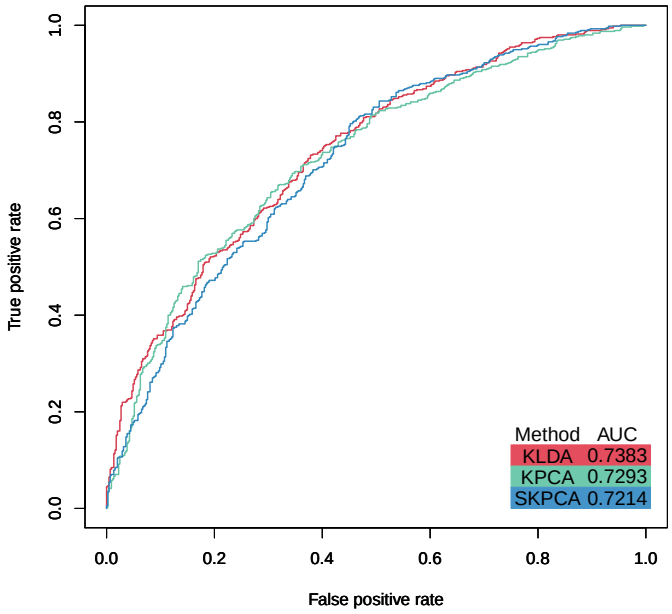


Figure 6. Receiver operating characteristic (ROC) curve and area under the curve (AUC) are compared by method for gender classification on FG-NET. Each color corresponds to a DR method.

For each fold, we transform and reduce the dimension of the features through each DR method. In all cases, a dimension of 100 is retained to facilitate comparison with the results on Morph-II. The transformed, dimension-reduced features then predict the gender of the testing fold's images through a linear SVM. The predicted classes from SVM are also mapped to probabilities through [78], similarly as in Section 6. The gender classification accuracy is calculated for the testing fold. Finally, all such

testing classification accuracies are averaged to compute the mean classification accuracy from testing; the testing probabilities are used to form ROC curves.

The optimum gender classification results on FG-NET are presented in Table 8. The maximum classification accuracy of about 72.25% is achieved by KLDA. For other choices of parameters, KLDA reaches above 71% accuracy, which is close to the maximum accuracy attained by SKPCA. Meanwhile, the peak accuracy reached by KPCA is 70.25%. In general here, KLDA is observed to outperform SKPCA and KPCA, while SKPCA tends to surpass KPCA. In most cases, the probability of correctly classifying males (sensitivity/true positive rate) is higher than the probability of correctly classifying females (specificity/true negative rate). For each DR method, an ROC curve (corresponding to the results from Table 8 with maximal mean classification accuracy) is displayed in Figure 6. The area under the curve (AUC) is highest for KLDA, followed by KPCA then SKPCA.

Overall, the gender classification results on Morph-II are stronger than on FG-NET. Lower accuracy on FG-NET could be caused by the greater number of minors (aged 0-18), who have been more difficult to classify than adults in some studies [35,82]. Additionally, there are substantially fewer faces for training in FG-NET versus Morph-II (under 1000 versus 10280 images). Another contributor could be the choice of features and its dimension; the AAM features have dimension 109 on FG-NET, while the HOG, LBP, and BIF features have dimensions ranging from 500 to thousands on Morph-II. SKPCA reaches peak performance on Morph-II, while KLDA attains optimal results on FG-NET. However, the results on Morph-II and FG-NET are similar in that the supervised methods KLDA and SKPCA outperform the unsupervised method KPCA for gender classification. Further, both datasets evidence that female faces are more challenging to classify than male faces.

8. Computational Framework for Practical Systems

To tackle the challenges of high dimensionality and intensive computation for large-scale databases (like Morph-II, as shown in the Time column of Table 5) in real-world applications, we propose a computational framework to substantially decrease runtime.

Our approach involves parallel computing, the bootstrap resampling method, and ensemble learning. Let $M1$ denote the main training set and $M2$ the testing set. If $M1$ is very large, we can save some time by drawing bootstrapped samples from $M1$. Let S_i denote the ith bootstrapped sample from $M1$. Send S_i to a core (or processor), Core i. Train the model on S_i. Test on the full testing set $M2$, obtaining a set of gender predictions corresponding to Core i and S_i. Repeat this process for all bootstrapped samples and corresponding cores i. The final predictions are obtained by taking the majority rule of the predictions from all i cores and samples. Hence, the results from this scheme approximate the results from the full Morph-II. This framework is summarized in Figure 7.

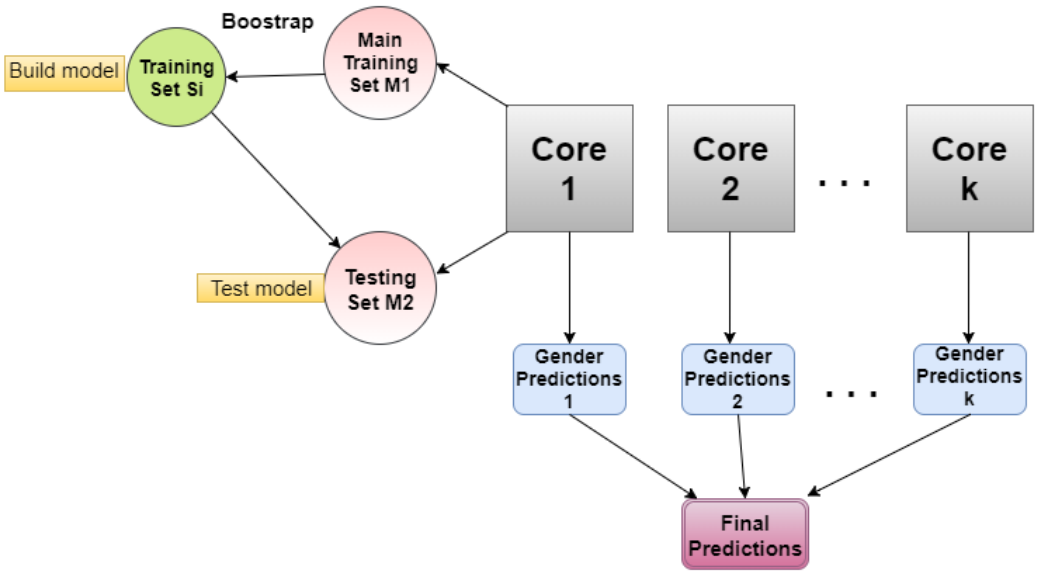


Figure 7. Flowchart representing the parallel computational framework for practical systems proposed for Morph-II and other large datasets.

To explore the effectiveness, this framework is applied to Morph-II with a selection of BIF, LBP, and HOG features as preliminary studies. This experiment is implemented through the HiPerGator 2.0 supercomputer at University of Florida with five cores per combination of feature and dimension reduction method. Following the subsetting scheme discussed in Section 5.2, for simplicity, we consider only the case of bootstrapping image samples from S_1 for training, while each image from $S_2 \cup R$ is used for testing.

Table 9. Classification Results Based on Bootstrapping

Method	Feature	Accuracy	Memory (gb)	Time (min)
KPCA	BIF $s = s7 - 37, \gamma = 0.4$	**0.9330**	27.59	90
	HOG $s = 12, o = 8$	0.9178	29.77	101
	LBP $s = 10, r = 1$	0.8927	25.85	37
SKPCA	BIF $s = s7 - 37, \gamma = 0.4$	**0.9417**	53.28	89
	HOG $s = 12, o = 8$	0.9056	51.43	74
	LBP $s = 10, r = 1$	0.9274	20.33	24
KLDA	BIF $s = s7 - 37, \gamma = 0.4$	**0.9416**	30.99	100
	HOG $s = 12, o = 8$	0.9133	25.42	102
	LBP $s = 10, r = 1$	0.9118	17.05	26

We evaluate this framework by comparing the approximated results in Table 9 to the results from Table 5. For each combination of feature and dimension reduction method, each of the five cores independently trains a bootstrapped sample of 1000 images from S_1 and tests on $S_2 \cup R$. Then the gender predictions over all five cores are compared with a simple majority rule; e.g., if an image is predicted male for three images and female for two images, the final gender prediction is male. The times in Table 9 are the total runtimes for this process, which include training and testing on HPC. Therefore, the times and memory can be compared between Tables 5 and 9. A distinction is that in Table 5, results are averaged for the alternating scheme, while in Table 9, the results are only from when S_1 is used for training and $S_2 \cup R$ for testing.

It is shown in Table 9 that, in many cases, the accuracy rates from the approximations are similar to those from the main approach in Table 5. This is a very good result, especially considering that the bootstrapping approach uses no more than 5000 images total for training, while the main approach used all 10280 images for training. This finding suggests that our methods may perform reasonably well on Morph-II with smaller training sets. The most substantial difference between the

bootstrapped approach and the main approach is in the runtime. For all combinations of features and dimension reduction methods, the bootstrapping approach has decreased the runtime to under two hours. Meanwhile, the main approach in Table 5 yields runtimes exceeding 20 hours. Hence, our preliminary results indicate the parallel approximation approach can attain similar accuracy rates to the main approach, while substantially saving time. Such a result is promising for practical gender classification systems, where gender predictions must be made in real-time.

9. Conclusion

We have performed a comparative study of the nonlinear dimension reduction methods KPCA, SKPCA, and KLDA. These kernel-based methods are first applied to three simulated datasets for visualization and comparison. SKPCA and KLDA outperform KPCA, reinforcing the need for supervised approaches in classification tasks. The radial kernel performed well, encouraging its use for face analysis.

Next, we have proposed and evaluated a new machine learning process for Morph-II. First, we use a novel subsetting scheme that reduces class imbalances while establishing independence between training and testing sets. Then we preprocess Morph-II photographs and extract three appearance-based features: HOG, LBP, and BIF. We transform and reduce the dimension of these features through KPCA, SKPCA, and KLDA. Linear SVM classifies the gender of Morph-II subjects, reaching accuracy rates of 95%. With promising preliminary results on Morph-II, a practical computational framework is offered that reduces runtime through parallelization and approximation.

The performance of the dimension reduction methods are further compared through an application to the FG-NET dataset. Images are represented through the appearance-based AAM features; transformed and reduced in dimension through KPCA, SKPCA, and KLDA; and classified as containing a male or female subject through linear SVM. While SKPCA performed optimally on Morph-II, KLDA reached top performance on FG-NET with 72% leave-one-person-out (LOPO) accuracy.

Further directions of research involve automatic tuning parameter selection, reduction of computational cost, and application to other face analysis tasks. Our approach could yield improved results with better choices of parameters, but it is impossible to anticipate and try all combinations. Automatic parameter selection for kernels could help identify a good set of parameters more easily. Perhaps the most important future direction of research on Morph-II is to reduce computational cost. For many practical demographic analysis systems, predictions must be made in real-time. For our gender classification methods, our parallel approximation approach substantially reduced runtime while attaining similar accuracy rates to the main approach. Such computational strategies should be further investigated to help bring gender classification and other face analysis tasks to practical implementation. Finally, our machine learning pipeline for Morph-II could be generalized to race classification or even age estimation.

10. Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant Numbers DMS-1659288. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the reviewers for the helpful comments that significantly improves the presentation of the paper.

1. Fodor, I.K. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.
2. Sorzano, C.O.S.; Vargas, J.; Montano, A.P. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877* **2014**.

3. Izenman, A.J. Modern multivariate statistical techniques. *Regression, classification and manifold learning* **2008**.
4. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572.
5. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **1933**, *24*, 417.
6. Yang, J.; Peng, W.; Ward, M.O.; Rundensteiner, E.A. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714). IEEE, 2003, pp. 105–112.
7. Johansson, S.; Johansson, J. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE transactions on visualization and computer graphics* **2009**, *15*, 993–1000.
8. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Annals of human genetics* **1936**, *7*, 179–188.
9. Rao, C.R. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)* **1948**, *10*, 159–203.
10. Lee, J.A.; Verleysen, M. *Nonlinear dimensionality reduction*; Springer Science & Business Media, 2007.
11. Nhan Duong, C.; Luu, K.; Gia Quach, K.; Bui, T.D. Beyond principal components: Deep boltzmann machines for face modeling. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4786–4794.
12. Yin, J.; Liu, Z.; Jin, Z.; Yang, W. Kernel sparse representation based classification. *Neurocomputing* **2012**, *77*, 120–128.
13. Shawe-Taylor, J.; Cristianini, N. *Kernel methods for pattern analysis*; Cambridge university press, 2004.
14. Motai, Y. Kernel association for classification and prediction: A survey. *IEEE transactions on neural networks and learning systems* **2015**, *26*, 208–223.
15. Xie, Y.; Luu, K.; Savvides, M. A robust approach to facial ethnicity classification on large scale face databases. Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on. IEEE, 2012, pp. 143–149.
16. Schölkopf, B.; Smola, A.; Müller, K.R. Kernel principal component analysis. International Conference on Artificial Neural Networks. Springer, 1997, pp. 583–588.
17. Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; Mullers, K.R. Fisher discriminant analysis with kernels. Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop. Ieee, 1999, pp. 41–48.
18. Baudat, G.; Anouar, F. Generalized discriminant analysis using a kernel approach. *Neural computation* **2000**, *12*, 2385–2404.
19. Barzilay, O.; Brailovsky, V.L. On domain knowledge and feature selection using a support vector machine. *Pattern Recognition Letters* **1999**, *20*, 475–484.
20. Schölkopf, B.; Simard, P.; Smola, A.J.; Vapnik, V. Prior knowledge in support vector kernels. Advances in neural information processing systems, 1998, pp. 640–646.
21. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *science* **2006**, *313*, 504–507.
22. Zhao, W.; Krishnaswamy, A.; Chellappa, R.; Swets, D.L.; Weng, J. Discriminant analysis of principal components for face recognition. In *Face Recognition*; Springer, 1998; pp. 73–85.
23. Martínez, A.M.; Kak, A.C. Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence* **2001**, *23*, 228–233.
24. Yang, J.; Yang, J.y. Why can LDA be performed in PCA transformed space? *Pattern recognition* **2003**, *36*, 563–566.
25. Turk, M.; Pentland, A. Eigenfaces for recognition. *Journal of cognitive neuroscience* **1991**, *3*, 71–86.
26. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence* **1997**, *19*, 711–720.
27. Kim, K.I.; Jung, K.; Kim, H.J. Face recognition using kernel principal component analysis. *IEEE signal processing letters* **2002**, *9*, 40–42.
28. Lu, J.; Plataniotis, K.N.; Venetsanopoulos, A.N. Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks* **2003**, *14*, 117–126.

29. Yang, J.; Zhang, D.; Frangi, A.F.; Yang, J.y. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE transactions on pattern analysis and machine intelligence* **2004**, *26*, 131–137.
30. Li, M.; Yuan, B. 2D-LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters* **2005**, *26*, 527–532.
31. Karg, M.; Jenke, R.; Seiberl, W.; Kühnlenz, K.; Schwirtz, A.; Buss, M. A comparison of PCA, KPCA and LDA for feature extraction to recognize affect in gait kinematics. Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on. IEEE, 2009, pp. 1–6.
32. Ye, F.; Shi, Z.; Shi, Z. A comparative study of PCA, LDA and Kernel LDA for image classification. Ubiquitous Virtual Reality, 2009. ISUVR'09. International Symposium on. IEEE, 2009, pp. 51–54.
33. Yang, J.; Jin, Z.; Yang, J.y.; Zhang, D.; Frangi, A.F. Essence of kernel Fisher discriminant: KPCA plus LDA. *Pattern Recognition* **2004**, *37*, 2097–2100.
34. Barshan, E.; Ghodsi, A.; Azimifar, Z.; Jahromi, M.Z. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition* **2011**, *44*, 1357–1371.
35. Wang, Y.; Chen, C.; Watkins, V.; Ricanek, K. Modified Supervised Kernel PCA for Gender Classification. International Conference on Intelligent Science and Big Data Engineering. Springer, 2015, pp. 60–71.
36. Fewzee, P.; Karray, F. Dimensionality reduction for emotional speech recognition. Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, 2012, pp. 532–537.
37. Samadani, A.A.; Ghodsi, A.; Kulić, D. Discriminative functional analysis of human movements. *Pattern Recognition Letters* **2013**, *34*, 1829–1839.
38. Wu, H.; Bowers, D.M.; Huynh, T.T.; Souvenir, R. Biomedical video denoising using supervised manifold learning. Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on. IEEE, 2013, pp. 1244–1247.
39. Ashtiani, H.; Ghodsi, A. A dimension-independent generalization bound for kernel supervised principal component analysis. Feature Extraction: Modern Questions and Challenges, 2015, pp. 19–29.
40. Fu, Y.; Guo, G.; Huang, T.S. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2010**, *32*, 1955–1976.
41. Sun, Y.; Zhang, M.; Sun, Z.; Tan, T. Demographic analysis from biometric data: Achievements, challenges, and new frontiers. *IEEE transactions on pattern analysis and machine intelligence* **2018**, *40*, 332–351.
42. Burton, A.M.; Bruce, V.; Dench, N. What's the difference between men and women? Evidence from facial measurement. *Perception* **1993**, *22*, 153–176.
43. Ng, C.B.; Tay, Y.H.; Goi, B.M. Vision-based human gender recognition: A survey. *arXiv preprint arXiv:1204.1611* **2012**.
44. Golomb, B.A.; Lawrence, D.T.; Sejnowski, T.J. Sexnet: A neural network identifies sex from human faces. NIPS, 1990, Vol. 1, p. 2.
45. Cottrell, G.W.; Metcalfe, J. EMPATH: Face, emotion, and gender recognition using holons. Advances in neural information processing systems, 1991, pp. 564–571.
46. Poggio, B.; Brunelli, R.; Poggio, T. HyberBF networks for gender classification, 1992.
47. Wiskott, L.; Fellous, J.M.; Krüger, N.; Von der Malsburg, C. Face recognition and gender determination, 1995.
48. Guo, G.; Mu, G. Human age estimation: What is the influence across race and gender? Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010, pp. 71–78.
49. Guo, G.; Mu, G. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. Computer vision and pattern recognition (cvpr), 2011 ieee conference on. IEEE, 2011, pp. 657–664.
50. Shan, C. Learning local binary patterns for gender classification on real-world face images. *Pattern recognition letters* **2012**, *33*, 431–437.
51. Yang, H.F.; Lin, B.Y.; Chang, K.Y.; Chen, C.S. Automatic age estimation from face images via deep ranking. *networks* **2013**, *35*, 1872–1886.

52. Yi, D.; Lei, Z.; Li, S.Z. Age estimation by multi-scale convolutional network. Asian conference on computer vision. Springer, 2014, pp. 144–158.
53. Antipov, G.; Berrani, S.A.; Dugelay, J.L. Minimalistic CNN-based ensemble model for gender prediction from face images. *Pattern recognition letters* **2016**, *70*, 59–65.
54. Antipov, G.; Baccouche, M.; Berrani, S.A.; Dugelay, J.L. Effective training of convolutional neural networks for face-based gender and age prediction. *Pattern Recognition* **2017**, *72*, 15–26.
55. Yang, Z.; Ai, H. Demographic classification with local binary patterns. International Conference on Biometrics. Springer, 2007, pp. 464–473.
56. Lian, H.C.; Lu, B.L. Multi-view gender classification using local binary patterns and support vector machines. International Symposium on Neural Networks. Springer, 2006, pp. 202–209.
57. Mäkinen, E.; Raisamo, R. An experimental comparison of gender classification methods. *pattern recognition letters* **2008**, *29*, 1544–1556.
58. Alexandre, L.A. Gender recognition: A multiscale decision fusion approach. *Pattern recognition letters* **2010**, *31*, 1422–1427.
59. Xia, B.; Sun, H.; Lu, B.L. Multi-view gender classification based on local Gabor binary mapping pattern and support vector machines. Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008, pp. 3388–3395.
60. Guo, G.; Dyer, C.R.; Fu, Y.; Huang, T.S. Is gender recognition affected by age? Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, 2009, pp. 2032–2039.
61. Han, H.; Otto, C.; Liu, X.; Jain, A.K. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **2015**, pp. 1148–1161.
62. Ma, Y.; Fu, Y. *Manifold learning theory and applications*; CRC press, 2011.
63. Schölkopf, B.; Herbrich, R.; Smola, A.J. A generalized representer theorem. International conference on computational learning theory. Springer, 2001, pp. 416–426.
64. Wahba, G. *Spline models for observational data*; Vol. 59, Siam, 1990.
65. Schölkopf, B.; Smola, A.; Müller, K.R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* **1998**, *10*, 1299–1319.
66. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. International conference on algorithmic learning theory. Springer, 2005, pp. 63–77.
67. Steinwart, I.; Christmann, A. *Support vector machines*; Springer Science & Business Media, 2008.
68. Ricanek, K.; Tesafaye, T. Morph: A longitudinal image database of normal adult age-progression. Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on. IEEE, 2006, pp. 341–345.
69. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intelligent data analysis* **2002**, *6*, 429–449.
70. Guo, G.; Mu, G. A study of large-scale ethnicity estimation with gender and age variations. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, 2010, pp. 79–86.
71. Yip, B.; Bingham, G.; Kempfert, K.; Fabish, J.; Kling, T.; Chen, C.; Wang, Y. Preliminary Studies on a Large Face Database. *arXiv preprint arXiv:1811.06446* **2018**.
72. Yip, B.; Towner, R.; Kling, T.; Chen, C.; Wang, Y. Image Pre-processing Using OpenCV Library on MORPH-II Face Database. *arXiv preprint arXiv:1811.06934* **2018**.
73. Edwards, G.J.; Taylor, C.J.; Cootes, T.F. Interpreting face images using active appearance models. Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. IEEE, 1998, pp. 300–305.
74. Kling, T. Morph-II: Feature Vector Documentation: NSF-REU site at UNC Wilmington. http://libres.uncg.edu/ir/uncw/f/wangy2018-1.pdf, 2017.
75. Byun, H.; Lee, S.W. Applications of support vector machines for pattern recognition: A survey. In *Pattern recognition with support vector machines*; Springer, 2002; pp. 213–236.
76. Qiu, Y.; Mei, J.; Qiu, M.Y. Package 'rARPACK' **2016**.
77. Dimitriadou, E.; Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A. Misc Functions of the Department of Statistics (e1071), TU Wien. *R package version* **2005**, pp. 1–5.

78. Platt, J.; others. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **1999**, *10*, 61–74.
79. Guo, G.; Mu, G. A framework for joint estimation of age, gender and ethnicity on a large database. *Image and Vision Computing* **2014**, *32*, 761–770.
80. Panis, G.; Lanitis, A.; Tsapatsoulis, N.; Cootes, T.F. Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics* **2016**, *5*, 37–46.
81. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. European conference on computer vision. Springer, 1998, pp. 484–498.
82. Wang, Y.; Ricanek, K.; Chen, C.; Chang, Y. Gender classification from infants to seniors. 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2010, pp. 1–6.