Abstract. High dimensionality, which is associated with a large number of predictors and a small sample size \( p \gg n \), and non-linearity are two main challenges of big datasets which can hinder the development of accurate statistical models. One field that would benefit from advances in dimensionality reduction (DR) is computer vision. This technology is the basis for human computer interaction, which has important applications in cybersecurity, medical diagnosis, and advertising. In this study we applied three non-linear DR algorithms: Kernel Fisher Discriminant Analysis (KFDA); Kernel Principal Component Analysis (KPCA); and Supervised Kernel Principal Component Analysis (SKPCA), to a large face image dataset, Morph-II. After DR, we conducted gender classification using a linear support vector machine (SVM-L) for comparative purposes in terms of prediction accuracy and computational cost.

1. Introduction

The kernel trick is widely used to facilitate the task of classification on a non-linearly separable dataset. This is accomplished by projecting the data to a subspace of arbitrarily high dimensionality, and exploring non-linear relationships among the features. Once a linearly separable representation is identified, we are able to perform the familiar linear form of a DR method, such as PCA or FDA, in the high-dimensional space. There are many well-known kernels including sigmoid, Gaussian, and polynomial. It is notable that the linear forms of PCA and FDA are special cases of the polynomial kernel, and so the linear form of these DR methods is a special case of the kernel version.

It is not always the case that one algorithm will outperform all others across a variety of datasets. In other words, different datasets pose different challenges, which might be handled better by one algorithm or another. Therefore, a data-driven approach, coupled with visualization, is an effective way to evaluate the performance of a set of algorithms on a given dataset. This approach is taken in the current study to evaluate the effectiveness of KPCA, KFDA, and SKPCA to separate six multi-class simulated datasets as well as subsets of the MORPH-II longitudinal face image database. The Morph-II database contains 55,134 mugshots of 13,617 individuals collected over 5 years. Subjects ages range from 16 to 77 years and while there is an average of around 4 pictures per individual, it ranges from 1 to 53 images per person. Metadata included with the database provide race, gender, date of birth, date of arrest, age, age difference since last picture, subject identifier, and id number for each image in the database.

2. Methodology

2.1. KPCA. In the following section we formulate KPCA as described in [?]. In linear PCA, the directions of highest variation in an \( n \times p \) dataset are identified by computing the eigenvectors which correspond to the largest eigenvalues of the \( p \times p \) dimensional covariance matrix of the centered data

\[
C = \frac{1}{n} \sum_{j=1}^{n} x_j x_j^T. \tag{1}
\]

where \( x_j \) is the centered observation vector corresponding to the \( j^{th} \) subject where \( j = 1, \cdots, n \). This is accomplished by solving the following eigenvalue equation,

\[
\lambda v = Cv, \tag{2}
\]

for \( \lambda > 0 \) and \( v \in \mathbb{R}^p \setminus \{0\} \).
In KPCA[7], this computation is performed in another inner product space which is related to the original space by a non-linear map $\Phi$, which maps the elements of $\mathbb{R}^p$ to a possibly infinite dimensional space $F$. This is done using the kernel trick, which involves replacing all occurrences of the inner product in the PCA algorithm by a kernel. In this work, the following Gaussian kernel is used:

$$K(x, x') = \exp(-\gamma||x - x'||^2),$$

(3)

Other possible choices include but are not limited to: the polynomial and sigmoid kernels. By using the kernel trick, it is not necessary to discover $\Phi$ explicitly, which greatly reduces the computational cost.

KPCA is an unsupervised DR method, meaning that it does not take into account class labels. Though the algorithm does find the largest sources of variation in the data, it is not necessarily associated with response of interest. The KPCA algorithm is as follows:

1. Compute the kernel matrix, $K_{ij} = (k(x_i, x_j))_{ij}$ for $i, j = 1, \ldots, n$.
2. Solve $n\lambda \alpha = K\alpha$ by diagonalizing $K$, where $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$ are the eigenvalues with corresponding eigenvectors $\alpha_1, \ldots, \alpha_n$.
3. Normalize eigenvectors $\alpha_j$ by requiring $\alpha_j^T \alpha_j = 1$, for $j = 1, \ldots, l$ where $\lambda_l$ corresponds to the smallest non-zero eigenvalue and $j = 1, \cdots, l$. Let $v_j$ be the normalized eigenvectors in the high-dimensional feature space.
4. Compute the projections of training data on to normalized eigenvectors $v_j$ using $v_j^T \Phi(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$.

2.2. SKPCA. In the following section we formulate SKPCA as described in [7]. SKPCA is similar to KPCA, except that class labels are used to maximize the dependency of the covariates on the response in question. This is done by using an estimate of the Hilbert Schmidt Independence Criterion,

$$HSIC(Z, F, G) = \frac{1}{(n-1)^2} tr(KHLH),$$

(4)

which measures the strength of association between two random variables, where $H, K, L \in \mathbb{R}^{n \times n}$, $K_{ij} = k(x_i, x_j)$, $L_{ij} = l(y_i, y_j)$, and $H_{ij} = [1(i = j) - 1/n]$, where $1(i = j)$ is an indicator function which is equal to $1$ when $i = j$ and 0 otherwise. In SKPCA, we seek to maximize the HSIC by maximizing:

$$tr(HKLH) = tr(HX^T U U^T XHLH) = tr(U^T XHLHXU).$$

(5)

Formulated as a constrained optimization problem we have

$$\text{argmax}_U \; tr(U^T XHLHXU)$$

subject to $U^T U = I$, which is solved by finding the eigenvectors of $Q = XHLHX^T$. By replacing all occurrences of $X$ by $\Phi(X)$, where $\Phi : \mathbb{R}^n \rightarrow F$ is a non-linear map to a high-dimensional space $F$, we arrive at the constrained optimization problem:

$$\text{argmax}_U \; tr(U^T \Phi(X) HHLH \Phi(X)^T U).$$

(6)

(7)

Using the fact that $U$ is a linear combination of the projected data, $U = \Phi(X) \beta$, we express the objective function as

$$tr(\beta^T \Phi(X)^T \Phi(X) HHLH \Phi(X)^T \Phi(X) \beta) = tr(\beta^T KHLHK \beta),$$

subject to $U^T U = \beta^T \Phi(X)^T \Phi(X) \beta = \beta^T K \beta$ where $K$ is the kernel matrix. The final form of our optimization problem is

$$\text{argmax}_\beta \; tr(\beta^T KHLHK \beta),$$

subject to $\beta^T K \beta = I$. The SKPCA algorithm is as follows:

1. Compute the kernel matrix, link matrix, and hat matrix, $K_{ij} = (k(x_i, x_j))_{ij}$, $L_{ij} = l(y_i, y_j)$, and $H_{ij} = [1(i = j) - 1/n]$ respectively.
2. Compute $Q = KHLHK$.
3. Find the new basis: $\beta$, the generalized eigenvectors of $(Q, K)$ corresponding to the top $d$ eigenvalues.
4. Project training data to new basis: $Z = \beta^T \Phi(X)^T \Phi(X) = \beta^T K$.
5. Project test data to new basis: $z = \beta^T \Phi(x)^T \Phi(x) = \beta^T K_{\text{test}}$.

Here, $K_{\text{test}}$ is the kernel matrix of the testing data, $X$ is the training data, $x$ is the testing data, $Z$ is the reduced dimension training data, and $z$ is the reduced dimension testing data.
2.3. **KFDA.** In the following section we formulate KFDA as described in [?]. KFDA is a non-linearization of FDA, which approximates the theoretically optimal Bayes’ classifier. FDA is a supervised method and so makes use of the classes to which the training data pertain to ensure association with the response of interest. In FDA we find the vector $w$, which maximizes the following objective function,

$$ J(w) = \frac{w^T S_B w}{w^T S_W w}, $$

where

$$ S_B = (\mu_c - \bar{x})(\mu_c - \bar{x})^T, \text{ and} $$

$$ S_W = \sum_c \sum_{i \in C} (x_i - \mu_c)(x_i - \mu_c)^T, $$

are the between-class and within-class scatter matrices respectively. $\bar{x}$ and $\mu_c$ are the global and within class means[?].

By replacing all occurrences of the inner product by the kernel function and using the fact that $w = \sum_i \alpha_i \Phi(x_i)$, we arrive at the non-linear formulation KFDA. KFDA focuses on maximizing the following objective function:

$$ J(\alpha) = \alpha^T S_B^\Phi \alpha $$

$$ \alpha^T S_W^\Phi \alpha, $$

where

$$ S_B^\Phi = \sum_c N_c[\kappa_c \kappa_c^T - \kappa \kappa^T], $$

$$ S_W^\Phi = KK^t - \sum_c N_c \kappa_c \kappa_c^T, $$

$$ \kappa_c = \frac{1}{N_c} \sum_{i \in c} K_{ij}, \text{ and} $$

$$ \kappa = \frac{1}{N} \sum_i K_{ij}, $$

are the kernel formulations of the between class scatter matrix, within class scatter matrix, within class mean, and global mean respectively. To solve for $\alpha$, we find the leading eigenvector of $N^{-1} M$, which maximizes the ratio of between class variation to within class variation in a high-dimensional feature space. Finally, we project new test points to the solution space using

$$ w^T \Phi(x) = \sum_i \alpha_i K(x_i, x). $$

3. **Simulation Studies**

Application of statistical methods to simulated datasets is a common preliminary step in data science research. We are able to engineer simulated data that pose specific geometric challenges which are exhibited by real data. In this work, PCA, FDA, KPCA, KFDA, and SKPCA were applied to three non-linearly separable, synthesized datasets, using the R statistical programming language. After dimension reduction, the data were projected to a low-dimensional space to gain insight through visual analysis. The R code, provided by Dr. Y. Wang of University of North Carolina at Wilmington, for KFDA, KPCA, and SKPCA. The generation of the following datasets can be found in the appendix. See Figures 1-3 for results.
Figure 1 contains simulated swiss roll data containing observations of four classes. KPCA failed to linearly separate the data, while KFDA and SKPCA were able to. Only KFDA was able to linearly separate the segmented concentric ring data in 2-dimensions. Referring to Figure 2, the wine chocolate data consists of a concentric sphere and outer shell, each of which containing observations from unique classes. All three DR methods were able to linearly separate the wine chocolate data. Figure 3 contains concentric segmented ring data from five classes. The linear DR methods were ineffective at linearly separating these three challenging datasets. Only KFDA and SKPCA were able to linearly separate this dataset. It seems that KFDA outperformed the other algorithms on all datasets. In general, KPCA and SKPCA were more sensitive to adjustments to the kernel tuning parameter. The tuning parameters, found using grid search, are not necessarily optimal. Optimal tuning parameters for KPCA and SKPCA would likely produce linearly separable projections with efficacy comparable to KFDA.

Figure 1. a) Simulated wine chocolate data in 3-dimensions. Different colors are used to denote data instances from different classes. b) Standard PCA is only able to rotate and reflect the data and so cannot create linearly separable classes. c) FDA was unable to linearly separate the classes. It projects the 2 class problem to the number line. d) Linearly separable wine chocolate data projected to 2-dimensions after application of the KFDA DR algorithm. e) Linearly separable wine chocolate data projected to 2-dimensions after application of the KPCA DR algorithm. f) Linearly separable wine chocolate data projected to 2-dimensions after application of the SKPCA DR algorithm.
Figure 2. a) Simulated swiss roll data in 3-dimensions. Different colors are used to denote data instances from different classes. b) Standard PCA is unable to linearly separate classes. c) Standard FDA is unable to linearly separate classes. d) Linearly separable swiss roll data projected to 2-dimensions after application of the KFDA DR algorithm. e) KPCA failed to linearly separate swiss roll data when projected to 2-dimensions. f) Linearly separable swiss roll data projected to 2-dimensions after application of the SKPCA DR algorithm.
4. APPLICATION TO FACIAL IMAGE DATA

Because of its size, its longitudinal span, and its inclusion of relevant metadata, MORPH-II has been used for a variety of race, gender, and age face imaging tasks. MORPH-II is one of the most popular face imaging databases available to the public\[?\]. Since its first release in 2006, it has been cited by over 500 researchers, as determined by our Google Scholar search. Multiple versions of MORPH have been released, but for our subsetting scheme we use the 2008 MORPH-II non-commercial release. MORPH-II includes over 55,000 mugshots with longitudinal spans from a few months to over twenty years. For each image, the following metadata is included: subject ID number, picture number, date of birth, date of arrest, race, gender, age, time since last arrest, and image filename.

Due to the high computational cost of kernel-based DR methods, subsets of 1000 random histogram-equalized images were selected from MORPH-II for analysis. Each subject has a single image in the subset. The ratio of male to female is 3:1 and black to white is 1:1. We applied the previously described DR methods (SKPCA still to be conducted, will be in final draft) to features extracted (cite Troy Kling webpage) using two methods, namely local binary patterns (LBP) and biologically inspired features (BIF). For a detailed discussion on the aforementioned feature extraction techniques see \[?\] and \[?\].
For LBPs, there are two main tuning parameters, radius and block size. We used LBPs with radii of 1,2, and 3 each with block sizes of 10, 12, 14, 16, 18, and 20. The BIFs have two main tuning parameters, gamma and block size. We used BIFs with gamma values of 0.1-1.0 incremented by 0.1 and block sizes of 7-37 and 15-29. We used SVM-L with 5-fold cross-validation to determine gender classification, standard deviation, and computational time using the R programming language.

### Table 1. Gender Classification Results

<table>
<thead>
<tr>
<th>Parameters</th>
<th>DR Method</th>
<th>Highest Accuracy</th>
<th>Standard Deviation</th>
<th>Computational Time (min.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBP $r=1,2,3$ $s=10,12,14,16,18,20$</td>
<td>KFDA</td>
<td>83.0% ($r=1$ $s=10$)</td>
<td>0.027</td>
<td>2.92</td>
</tr>
<tr>
<td>KPCA</td>
<td>86.7% ($r=1$ $s=14$)</td>
<td>0.013</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>SKPCA</td>
<td>88.8% ($r=3$ $s=14$)</td>
<td>0.015</td>
<td>3.06</td>
<td></td>
</tr>
<tr>
<td>BIF $g=0.1,0.2,...,1.0$ $s=15-29,7-37$</td>
<td>KFDA</td>
<td>83.3% ($g=1.0$ $s=7-37$)</td>
<td>0.27</td>
<td>3.11</td>
</tr>
<tr>
<td>KPCA</td>
<td>90.8% ($g=0.2$ $s=7-37$)</td>
<td>0.019</td>
<td>2.33</td>
<td></td>
</tr>
<tr>
<td>SKPCA</td>
<td>91.9% ($g=1.0$ $s=7-37$)</td>
<td>0.156</td>
<td>2.78</td>
<td></td>
</tr>
</tbody>
</table>

5. Discussion/Conclusions

KFDA, KPCA, and SKPCA were applied to LBP feature sets with radius values 1,2,3 and block size values of 12-20 incremented by 2. The prediction accuracy was calculated using 5-fold cross-validation. The highest gender prediction accuracy for KFDA, 83.0%, resulted from using radius 1 and block size 10. The dimensionality was reduced from 2479 to 3. The highest gender prediction accuracy for KPCA, 86.7%, resulted from using radius 1 and block size 14. The dimensionality was reduced from 1181 to 3. The highest gender prediction accuracy for SKPCA, 88.8%, resulted from using radius 3 and block size 14. The dimensionality was reduced from 6076 to 499.

KFDA, KPCA, and SKPCA were also applied to BIF feature sets with gamma values of 0.1-1.0 incremented by 0.1 and block sizes of 15-29 and 7-37. The highest gender prediction accuracy for KFDA, 83.3%, resulted from using gamma set to 1.0 and block size 14. The dimensionality was reduced from 2569 to 3. The highest gender prediction accuracy for KPCA, 90.8%, occurred from using gamma set to 0.2 and block size 7-37. The dimensionality was reduced from 2569 to 3. The highest gender prediction accuracy for SKPCA, 91.9%, resulted from using gamma 1 and block size 7-37. The dimensionality was reduced from 2569 to 499.

6. Acknowledgements

A cordial thank you is due to Dr. Yishi Wang, Dr. Cuixian Chen, Troy Kling, and Katherine Kempfert for their tutelage and contributions to this work. Also, thank you to the National Science Foundation (Grant #: 1659288) for establishing this undergraduate research experience and providing funding.

### References


