

USING SPARSE REGRESSION TO LEARN EFFECTIVE PROJECTIONS FOR FACE RECOGNITION

Yongxin Taylor Xi and Peter J. Ramadge

Dept. Electrical Engineering, Princeton University, Princeton NJ 08544

ABSTRACT

We explore sparse regression for effective feature selection and classification in face identity and expression recognition. We argue that sparse regression in pixel space is inappropriate. We propose instead a method which combines the virtues of sparse regression with projection methods such as PCA and FDA. The method can learn a sparse set of discriminative projections and increase recognition accuracy beyond that achievable by FDA. We demonstrate this by performance comparisons on three face data sets.

Index Terms— Face recognition, Feature extraction, Image recognition, Pattern classification, Object detection.

1. INTRODUCTION

The important problems of face and face expression recognition from images have received considerable attention, see e.g. [1, 2, 3]. The problems are challenging for several reasons. First, the dimension $p = mn$ of the images is usually much larger than the number of examples N . So from few examples one must learn to discriminate in a high dimensional space. Second, the example images exhibit variation both in the dimension of interest, e.g. identity, and other nuisance dimensions, e.g. lighting, pose, expression, background, etc. Third, face images presumably lie on a low dimensional, non-linear manifold \mathcal{X} . However, the detailed structure of this manifold is unavailable. Hence face recognition must use relatively few labelled points on an unknown manifold in high dimensional space to discriminate face identity while being robust to nuisance variations.

The above challenges lead naturally to consideration of feature selection and dimensionality reduction to first identify, and then focus on, a relatively small set of informative variables. Principle Component Analysis (PCA) [4, 5] and Fisher Linear Discriminant Analysis (FDA) [6, 5], are common examples of this approach. PCA selects projections that maximize the variance of the projected data. This is designed to accurately approximate the data by capturing aspects of greatest variation. However, accurate approximation need not result in the selection of informative features. Indeed, a face recognition study in [5] found that the first three PCA projections were mainly associated with lighting conditions and

discarding these usually improved recognition performance. FDA, on the other hand, uses the label information to select projections that maximize class separability. Underlying FDA is the assumption that each class has a Gaussian distribution with common covariance but distinct mean. Hence FDA seeks to discriminate classes by projecting onto a direction w that maximizes the distance between projected means (the discriminating signal) while minimizing projected class variation (the noise). In practical face recognition problems, FDA performs better than PCA [5]. Both PCA and FDA also have various (robust, regularized, kernelized) extensions [7, 8, 2].

The key contribution of this paper is an alternative means of selecting informative features prior to classification. Specifically, we formulate the projection selection problem as a sparse linear regression over a set of admissible projections. The sparsity constraint promotes conservative projection selection - each selected projection incurs a cost that must be compensated by efficacy in classification. Sparse regression also allows the aggregation of projections obtained from various criteria, e.g. wavelet bases, PCA, local PCA and FDA. This may compensate for the deficiencies of any particular projection scheme. However, inherent in any linear regression scheme is an assumption, as in FDA, of linear separability. We show experimentally that the proposed scheme outperforms nominal FDA in practical face and expression recognition tasks. Moreover, our formulation appears more effective in feature selection than previous regression methods for face recognition [9, 10].

We provide additional background to the problem in §2. Then formulate the projection selection algorithm in §3 and present experimental results using this algorithm in §4. We summarize our conclusions in §5.

2. BACKGROUND

We assume that a $m \times n$ grey scale image is represented as a point in $x \in R^p$ with $p = mn$. We are given a set of N training examples $\{(x_i, y_i)\}_{i=1}^N$ with x_i the i th face instance and $y_i \in \{1, \dots, k\}$ its label. We assume the training instances are centered ($\sum_j x_j = 0$) and let $X = [x_1, x_2, \dots, x_N]^T$.

Our objective is to combine projections with sparse regression to learn a classifier $h: \mathcal{X} \rightarrow Y$ that classifies new face images well. We first give some additional background

on PCA, FDA and regression classification. PCA finds the ON eigenvectors w_j , $j = 1, \dots, N - 1$, ordered by eigenvalue, largest to smallest, of the scatter matrix

$$X^T X = \sum_{k=1}^N x_k x_k^T \in \mathbf{R}^{p \times p}$$

Let $P_{PCA} = [w_1, w_2, \dots, w_{N-1}]$. Projecting an instance x onto the first d eigenvectors yields projected instance $\hat{x} \in \mathbf{R}^d$.

By contrast, FDA seeks to maximize the ratio of between-class to within-class scatter:

$$\max_w \frac{w^T S_B w}{w^T S_W w}, \quad \text{subject to } w^T w = 1. \quad (1)$$

where $S_B, S_W \in \mathbf{R}^{p \times p}$ are the between-class and within-class scatter matrices:

$$S_B = \sum_{c=1}^k N_c \mu_c \mu_c^T$$

$$S_W = \sum_{c=1}^k \sum_{x_i \in \text{class } c} (x_i - \mu_c)(x_i - \mu_c)^T$$

N_c is the number, and $\mu_c \in \mathbf{R}^p$ the mean, of instances in class c . When $N < p$, S_W is ill-conditioned and is replaced by the regularized estimator $S_W + \rho I$, small $\rho > 0$. Problem (1) then reduces to the eigenvalue problem $(S_W + \rho I)^{-1} S_B w = \lambda w$. There are at most $k - 1$ non-negative eigenvalues $\lambda_1 > \dots > \lambda_{k-1}$ (S_B has rank $\leq k - 1$) with eigenvectors w_1, \dots, w_{k-1} . Let $P_{FDA} = [w_1, w_2, \dots, w_{k-1}]$. By restricting attention to the first $d \leq k - 1$ coordinates in any extension of P_{FDA} to a basis, we obtain a projection into \mathbf{R}^d .

Both PCA and FDA project instances to points $\hat{x} \in \mathbf{R}^d$, $d \ll p$. A simple nonlinear classifier can then be used in \mathbf{R}^d to label each projected instance, e.g. a nearest neighbor classifier using the labelled projected training instances.

In regression based classification, the discrete label j is encoded as the j th standard basis vector e_j in \mathbf{R}^k . Let $Y = [y_1, y_2, \dots, y_N]^T \in \mathbf{R}^{N \times k}$ denote this encoding of the training labels. We then seek regression coefficients $B \in \mathbf{R}^{p \times k}$ to minimize the squared loss:

$$\min_B \|XB - Y\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The matrix B^T maps instances into \mathbf{R}^k . A nearest neighbor classifier in \mathbf{R}^k can then be used to label: $h(x) = \operatorname{argmin}_{j=1}^k \|B^T x - e_j\|_2$. Column j of B defines a binary classifier for class j against the remaining $k - 1$ classes.

Usually $N < p$ and the regression problem requires regularization. The most common approach, ridge regression, considers

$$\min_B \|XB - Y\|_F^2 + \lambda \|B\|_F^2$$

See e.g. [9]. In general, the corresponding solution

$$B = (X^T X + \lambda I)^{-1} X^T Y$$

is not a sparse matrix. Hence it potentially uses all of the pixels in a face image to project and then classify.

3. PROJECTION VIA SPARSE REGRESSION

An alternative means of regularizing the regression classifier is to require that B be a sparse matrix. This encourages the identification of features most important for classification and has the potential for better generalization [11]. A computationally tractable means of achieving sparsity is by penalizing the l_1 norm of the regression coefficients, e.g. LASSO [12]. This results in the convex optimization problem:

$$\min_B \|XB - Y\|_F^2 + \lambda \sum_{i,j} |B_{i,j}|$$

This decomposes into k subproblems, one for each class:

$$\|Xb_i - y_i\|_F^2 + \lambda \|b_i\|_1, \quad \text{where } y_i = Y_{\text{col } i}, i = 1, \dots, k.$$

This formulation determines a sparse set of informative pixels for the classification problem. However, we do not expect face identity to be robustly captured by a sparse set of pixels.

Instead, we propose to regress over a set of admissible projections of pixel values. For example, these could be a subset of PCA projections, FDA projections, or some the union of these or other projections. Let the columns of $Q \in \mathbf{R}^{p \times m}$ be the set of admissible (linear) projections. Then we seek the solution $B = [b_1, \dots, b_k] \in \mathbf{R}^{m \times k}$ of:

$$\min_{b_j} \|XQb_j - y_j\|_F^2 \quad \text{s.t. } |b_j|_1 \leq r_j \quad j = 1, \dots, k \quad (2)$$

Notice that XQ becomes the new data matrix, yielding m regressors for b_j to sparsely weight. The parameters $r_j > 0$ control the sparsity of the b_j . The nonzero coefficients of b_j combine a few projections in Q to obtain informative projections for classification of class j against the remaining classes. We can infer the degree of importance of these projections by the corresponding weights. Finally, setting $Q = I$ reduces the problem to seeking a sparse set of informative pixels.

There are a variety of existing algorithms for solving (2). In §4 we employ Least Angle Regression (LARS) [13] to obtain an approximate solution.

4. EXPERIMENTS

We use three standard face databases, YALE, ORL and FERET [14, 15], to experimentally compare the proposed method with nominal implementations of PCA and FDA followed by nearest neighbor classification. For clarity, we only present results for the proposed method with $Q = P_{PCA} \in$

$\mathbf{R}^{p \times (N-1)}$ (PCA-LARS), $Q = P_{FDA} \in \mathbf{R}^{p \times (k-1)}$ (FDA-LARS), and $Q = I$ (I-LARS).

The YALE images (15 subjects, 11 images per subject, 320×243 image size) were first centered, histogram normalized and resized to 60×50 pixels. The ORL images (40 subjects, 10 images per subject, 112×92 image size) were used without preprocessing. A subset of FERET gray images (200 subjects, 7 images per subject, 384×256 image size) were centered, histogram normalized and resized to 120×100 pixels. The 7 images are frontal, frontal with expression, frontal with different illumination, $\pm 15^\circ$ and $\pm 25^\circ$ profiles. (corresponding to the b series in FERET database documentation: ba, bj, bk, be, bf, bd, bg.)

We repeated each experiment 20 times with random selection of the training set (per person: 2 and 6 for YALE, 5 for ORL, 4 for FERET). We tested on the remaining data and averaged the test error of the 20 runs. For each method we use a Euclidean nearest-neighbor classifier. The table below summarizes the key features of the data sets.

Data Set	$m \times n$	N	k	$mn/(N/k)$
YALE	60×50	30 or 90	15	1500 or 500
ORL	112×92	200	40	2060
FERET	120×100	800	200	3000

We first tested the different algorithms on YALE using 2 and 6 (out of 11) training images per person. See Fig. 1(m) & 1(n). Not surprisingly, PCA-LARS outperformed PCA in both cases. Indeed the performance of PCA-LARS compared well with that of FDA. This was particularly evident at small projection dimensions and for fewer examples per person. Perhaps more surprising was that FDA-LARS consistently outperformed FDA. This was significant for 2 examples per person, less so for 6 examples per person. For four of the classes, an example image and the first two LARS-selected PCA eigenfaces are shown in Fig. 1(a)-1(l). These are the PCA projections, as determined by sparse regression, that best discriminate the class from the remaining $k - 1$ classes. As predicted, selecting a sparse set of informative pixels (I-LARS) did not perform well.

Next, we studied the effect of the number of classes k . To do so we created subsets of FERET with $k = 10, 20, 40$, corresponding to the first 10, 20 and 40 subjects respectively. In this experiment we first used full PCA to reduce the dimension of the data to $N - 1$. Then we applied the various algorithms. Fig. 2 reports the resultant classification error. FDA-LARS always performed better than FDA but its advantage over FDA shrinks as k , and hence the difficulty of the task, increases.

In the third experiment we tested the algorithm's ability to recognize facial expressions and to detect eyeglasses. The eyeglass database is a subset of ORL. There are 30 examples for each class (eyeglasses, no eyeglasses), from which 20 examples were randomly selected for training. Fig. 3(m) shows

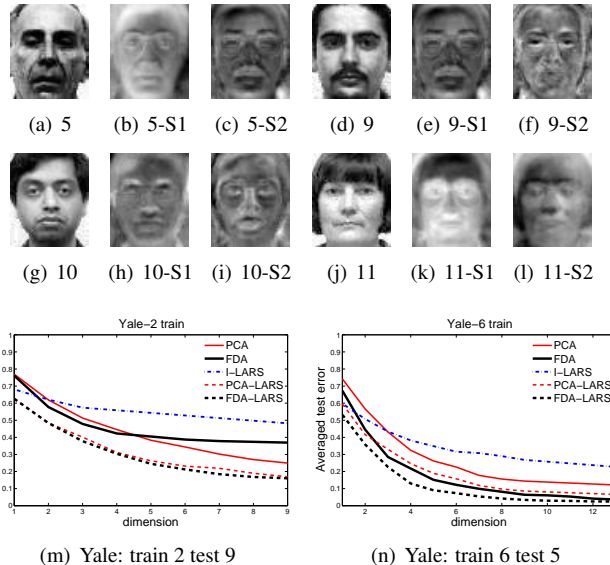


Fig. 1. Exp. 1: Results on YALE.

plots of the averaged detection accuracy over 20 runs. As expected, PCA-LARS outperformed PCA due to its ability to select informative eigenfaces. Indeed, the first 4 selected eigenfaces S1 to S4 (Fig. 3(c)-3(f)) subjectively appear to be discriminative, while the principal eigenfaces P1 and P2 (Fig. 3(a)-3(b)) do not. FDA had similar performance to PCA and FDA-LARS performance was the best.

Finally, we tested the algorithms on the happy and sad faces from YALE. For each subject, YALE has one example each of a happy and sad face, for a total of 30 instances. We randomly selected 10 training images for each expression (20 labelled instances) then tested on the remaining 10. The results are shown in Fig. 3(n). Images S1 to S4 (Fig. 3(i)-3(l)) are the eigenimages selected by PCA-LARS. Although a subjective evaluation, these images appear to exhibit expression information.

For the eyeglasses and expression experiments, running the sparse regression with $Q = I$, i.e., seeking a sparse set of informative pixels, performed reasonably well (better than PCA and FDA - results not shown). Reflecting that in these images eyeglasses and facial expressions are spatially localized in a small subset of informative pixels.

5. CONCLUSION

The sparse regression method empirically improved the recognition accuracy of both PCA and FDA projection classification. Since PCA does not use label information in projection selection, the improvement for PCA-LARS was to be expected. More surprising was that it consistently improved the recognition accuracy of FDA. Using sparse regression to select d of the $k - 1$ FDA projections yielded better per-

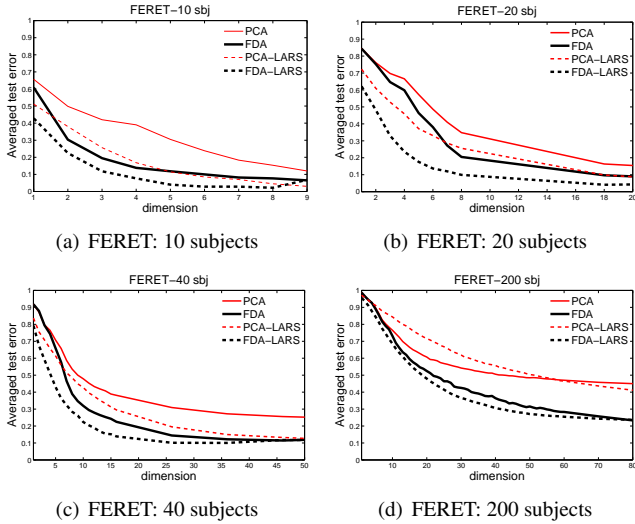


Fig. 2. Exp. 2: Results on FERET.

formance that simply selecting the first d FDA projections. The method performed particularly well in face expression classification and in eyeglasses recognition - far better than FDA. Since the method does not preselect a small set of projections (as in nominal PCA and FDA) it has the potential to discover discriminating projections that might be otherwise overlooked.

6. REFERENCES

- [1] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, pp. 399–458, 2003.
- [2] G. Shakhnarovich and B. Moghaddam, "Face recognition in subspaces," in *S.Z. Li, A.K. Jain (Eds.), Handbook of Face Recognition*. 2004, pp. 141–168, Springer.
- [3] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, May 1995.
- [4] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [5] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul 1997.
- [6] K. Etemad, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of America A*, vol. 14, pp. 1724–1733, 1997.
- [7] S. Kim, A. Magnani, and S. P. Boyd, "Robust fisher discriminant analysis," in *In Advances in Neural Information Processing Systems*. 2006, pp. 659–666, MIT Press.
- [8] J. Lu and et. al., "Regularization studies of linear discriminant analysis in small sample size scenarios with

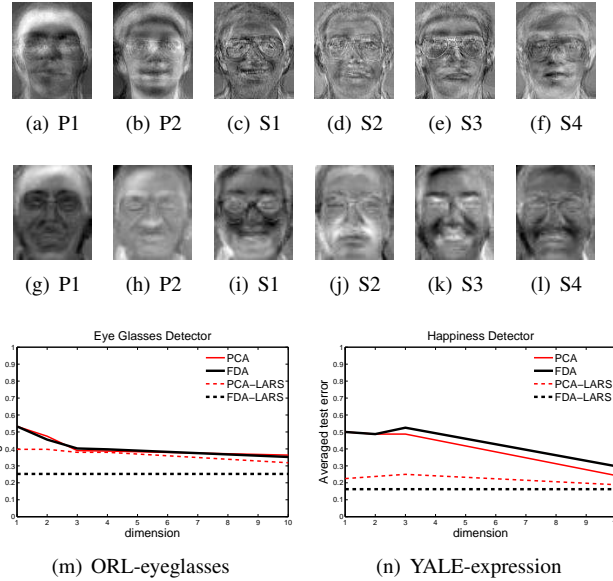


Fig. 3. Exp. 3: ORL-eyeglasses & YALE-expression.

application to face recognition," *Pattern Recogn. Lett.*, vol. 26, no. 2, pp. 181–191, 2005.

- [9] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," *CVPR '07*, pp. 1–7, 2007.
- [10] L. H. Clemmensen, D. D. Gomez, and B. K. Ersbøll, "Individual discriminative face recognition models based on subsets of features," in *SCIA 2007, LNCS 4522 proceedings*, 2007, pp. 61–71.
- [11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, August 2001.
- [12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.
- [13] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.
- [14] P. J. Phillips and et. al., "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, pp. 1090–1104, 2000.
- [15] P. J. Phillips and et. al., "The feret database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295 – 306, 1998.