Sparse Eigenfaces: The Effects of Elastic Net Regularization on the Principal Components of Facial Images

Sanandeesh Kamat[1]*, Sunish Shah[2], Yuanshuo Qu[3]

Abstract
Principal component analysis(PCA) is a classical tool in multivariate data analysis for dimensionality reduction. It uses vector space transformation to reduce high-dimensional data to lower dimensions. Specifically, it transforms the input variables into principal components that corresponds to directions of maximal variance in the dataset. Despite being widely used, ordinary PCA suffers from difficulty in interpretation of principal components, which are linear combinations of all the original variables. Sparse principal component analysis (SPCA) solved this problem by using LASSO(Elastic Net) to produce principal components with sparsity. In this project, we have applied both Ordinary PCA and SPCA with the extended Yale face dataset B, and the differences in PC-loading and PC-score separabilities were investigated. It was found that sparse PC-sores could separate images according to lighting conditions with greater distance than ordinary PC scores.

Keywords
Sparse PCA, Elastic Net, Eigenfaces

[1] *Department of Electrical & Computer Engineering, Rutgers University, NJ, USA*
[2] *Department of Statistics & Biostatistics, Rutgers University, NJ, USA*
[3] *Department of Genetics, Rutgers University, NJ, USA*
***Corresponding author**: ssk93@scarletmail.rutgers.edu

Contents

Introduction

PCA is widely used as an exploratory data analysis and predictive modelling tool. The operation can be conceptualized as a rotation of multi-dimensional input data points onto orthogonal axes of greatest relative variance (i.e. their Principal Components (PC)). Each PC in order of principality bears a lesser amount of variance, and together represent an n-dimensional ellipsoid fit onto the data. Hence, the omission of transformed features along the PCs of lowest principality results in minimal information loss. A high dimensional data set through which only a few prevailing forces are acting, can be projected down to lower dimensions where the effects of these forces are maximally retained. Dimensionality reduction of complex input data sets improves both analysts' intuition for the internal structure of the data, as well as the performance of predictive models which are otherwise susceptible to over-fitting noise.

The input data $\mathbf{X}$ is an $n \times p$ matrix for which there are n observed samples and p sample features. PCA can be performed either by Eigenvalue Decomposition (EVD) of the sample covariance matrix $\mathbf{X}^T\mathbf{X}$ or, more generally, by Singular Value Decomposition (SVD) of $\mathbf{X}$. Conventionally, each column of $\mathbf{X}$ is either zero centered or standardized. SVD states that $\mathbf{X}$ can be factorized as

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{1}$$

where $\mathbf{U}$ is the $n \times n$ matrix of *left singular vectors*, Σ is the $n \times p$ diagonal matrix of *singular values*, and $\mathbf{V}$ is the $p \times p$ matrix of *right singular vectors*. Whereas the orthonormal columns of $\mathbf{U}$ are the $n \times 1$ eigenvectors of $\mathbf{X}\mathbf{X}^T$, the orthonormal columns of $\mathbf{V}$ are the $p \times 1$ eigenvectors $\mathbf{X}^T\mathbf{X}$. The columns of $\mathbf{V}$ are referred to as the **PC-loadings** of $\mathbf{X}$. The PC loadings transform the rows of $\mathbf{X}$ from the original (possible correlated) feature-space into the desired orthogonalized PC-space. These

orthogonolized features $\mathbf{Z}$ are referred to as **PC-scores** of $\mathbf{X}$,

$$\begin{aligned} \mathbf{Z} &= \mathbf{XV} && (2)\\ &= \mathbf{U\Sigma V}^T\mathbf{V} && (3)\\ &= \mathbf{U\Sigma}, && (4) \end{aligned}$$

which sequentially capture the maximum variance between the features of $\mathbf{X}$. However, the uniformly non-zero PC loadings $\mathbf{V}$ ensure that all features of $\mathbf{X}$ contribute to $\mathbf{Z}$. Zou et al (2006) claim that PC loadings would benefit from sparsity in the same manner that multiple linear regression models do. Namely, that sparse model coefficients reduce overfitting and performance variance. Although naive thesholding (i.e. variable subset selection) has been applied in this setting, it is shown to produce potentially miseading results (Jollife 1995).

1. Applying Regularization to PCA

The same requirement for sparisty arises in multiple linear regression, for which predictions are ordinarily linear combinations of all p sample features. A promising balance of model accuracy and sparisty is achieved with regularization methods such as **lasso** and **elasticnet**. This section describes how these regression-based regularization techniques can be applied to the PCA, resulting Sparse PCA (**SPCA**) (Zou et al, 2006).

1.1 Lasso & Elasticnet Regularization

Consider that $\mathbf{X}$ is a zero-centered matrix of n observations and p predictors for a linear regression model, for which $\mathbf{Y}$ is the $n \times 1$ response vector we seek to fit. The ordinary least squares criterion for minimizing MSE (i.e. maximizing likelihood) is

$$\hat{\beta}_{OLS} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|^2 \tag{5}$$

Sparsity can be imposed upon β_{OLS} by introducing the lasso regularization criterion,

$$\hat{\beta}_{lasso} = \min_{\beta} \left(\|Y - \mathbf{X}\beta\|^2 + \lambda_l \|\beta\|_1 \right), \tag{6}$$

for which the lasso parameter $\lambda_l > 0$. In terms of Bayesian optimality, this criterion assumes a zero-mean laplacian prior distribution on β_{lasso}. Hence, the lasso penalty shrinks the regression coefficients towards zero; possibly exactly to zero. The resulting β_{lasso} estimates are both accurate and sparse, which are qualities sought for the SPCA loadings as well. However, the number of non-zero β_{lasso} is upper bounded by the number of samples. This can be unfavorable in situations where $n << p$, such as with image samples. Fortunately, the convex combination of lasso and ridge penalties into a single criterion retains the potential for all p features to be included in β. This combined criteria is referred to as elastic-net,

$$\hat{\beta}_{en} = \min_{\beta} \left(\|Y - \mathbf{X}\beta\|^2 + \lambda_l \|\beta\|_1 + \lambda_r \|\beta\|^2 \right), \tag{7}$$

for which lasso and ridge parameters $\lambda_{l,r} > 0$. For a fixed λ_r, the LARS-EN algorithm efficiently solves β_{en} across a range of λ_l (Zou and Hastie, 2005).

1.2 Sparse PCA

This section reframes PCA as a multiple linear regression problem, and then introduces a modified PCA based on the elastic-net criterion. Consider $\mathbf{Z}_i = \mathbf{U}_{,i}\Sigma_{i,i}$ to be the $n \times 1$ vector of ith PC-scores. Then, given $\mathbf{Z}_i$ and $\mathbf{X}$, the ith PC-loadings $\mathbf{V}_{,\mathbf{i}}$ can be approximatd by normalized $\hat{\beta}$ through

$$\hat{\beta} = \min_{\beta} \left(\|\mathbf{Z}_i - \mathbf{X}\beta\|^2 + \lambda_l \|\beta\|_1 + \lambda_r \|\beta\|^2 \right). \tag{8}$$

Although the positive ridge penalty $\lambda_r \|\beta\|^2$ ensures a unique solution when $p > n$ or $\mathbf{X}$ is rank deficient, it does not penalize the elements of β. Only the positive lasso penalty $\lambda_l \|\beta\|^2$ penalizes the elements of β, and, if λ_l is large enough, shrinks them to zero. We have that $\hat{V}_i = \frac{\hat{\beta}}{\|\hat{\beta}\|}$ approximates the ith PC-loading V_i and the transformation $\mathbf{X}\hat{V}_i$ approximates the ith PC-scores $\mathbf{Z}_i$. Unfortunately, the criterion stated in (8) requires first computing the original PC-scores, $\mathbf{Z}_i$. An alternative theorem is sought which is dependant solely on the input data, $\mathbf{X}$.

Consider two non-negative vectors $\alpha, \beta \in \Re^p$. We have that if

$$(\hat{\alpha}, \hat{\beta}) = \min_{\alpha,\beta} \sum_{i=1}^{n} \|x_i - \alpha\beta^T x_i\|^2 + \lambda_r \|\beta_j\|^2, \tag{9}$$

subject to $\|\alpha\|^2 = 1$, then $\hat{\beta}$ is proportional to the leading PC-loadings V_1 by a scale factor, i.e. $\hat{\beta} \propto V_1$.

To extend this theorem to the first k PC-loadings, suppose that $\mathbf{A}_{p\times k} = [\alpha_1 \cdots \alpha_k]$ and $\mathbf{B}_{p\times k} = [\beta_1 \cdots \beta_k]$. We have that if

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \min_{A,B} \sum_{i=1}^{n} \|x_i - AB^T x_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2, \tag{10}$$

subject to $A^T A = I_{k\times k}$, then $\hat{\beta}_j \propto V_j$ for $j = 1, \cdots k$. Finally, with the self-contained regression-based criterion for PCA in (10), sparsity to β is made possible by adding the lasso penalty,

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \min_{A,B} \sum_{i=1}^{n} \|x_i - AB^T x_i\|^2 + \lambda \sum_{j=1}^{k} \|\beta_j\|^2 + \sum_{j=1}^{k} \lambda_{1,j} \|\beta_j\|_1, \tag{11}$$

subject to $A^T A = I_{k\times k}$. The criterion (11) is henceforth referred to as the **SPCA criterion**. As expected, it reduces to original PCA when the lasso penalty vanishes, $\lambda_l = 0$.

2. Materials & Methods

This section describes the software resources used to conduct an experimental survey of SPCA. The goal was to apply PCA and SPCA to a large dataset of facial images and compare and cotrast the two sets of results. In particular, the differences in (1) PC-loading structure and (2) PC-Score seperabilities (wrt. sample metadata) were investigated.

2.1 The Extended Yale Face Database B

The *Extended Yale Face Database B* was produced by Georghiades et al. 2001. It consists of tightly cropped (192×168) pixel images of 38 subjects' faces. Each subject was photographed at 64 unique light source angles (i.e. azimuth and elevation). A scatter plot of these light source angles is provided in Figure 1.

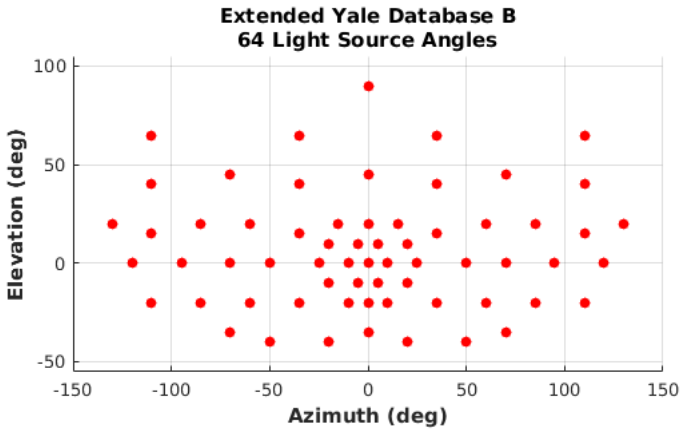

Figure 1. Light Source Azimuths and Elevations for Sample Images

To save on computation time, all $n = 2432$ images were downsampled by a factor of 4 (to $p = 8064$), before being stored as row vectors in a $n \times p$ input matrix, $\mathbf{X}$. The full input matrix and the average image are shown in Figure 2. The light source azimuths and elevations for each sample were stored in label matrix, $\mathbf{Y}_{n\times 2}$.

Because many of the image pixels may be unilluminated by light, or generally uninformative, it was expected that the latent structure of the samples are sparse. Therefore, it was expected that SPCA would eliminate unnecessary PC-loadings, without sacrificing seperability of meaningful factors (e.g. light source angles). Moreover, it was expected that SPCA would improve the interpretation of those PC-loadings which truly matter.

2.2 SpaSM: Sparse Statistical Modeling

The *SpaSM: A Matlab Toolbox for Sparse Statistical Modeling*, developed by Sjöstrand et al, 2010, was used to supply the SPCA software capabilities. This entire toolbox implements regularized regression, classification, and decomposition techniques based on the path-following paradigms developed at Stanford University; begining with (Efron et al, 2004). These techniques (e.g. Least Angle Regression (LAR), LASSO, Elastic Net, and SPCA) sequentially produce coefficients that are piecewise linear functions of the penalty parameter λ, and become sparse under suitable amounts of regularization. All of the software functions SpaSM have been validated against their respective publications.

Table 1, derived from Zou et al. 2006 and Sjöstrand et al, 2010, describes the sequential operations of the SPCA. Because this data set of images posesses $p >> n$, the ridge

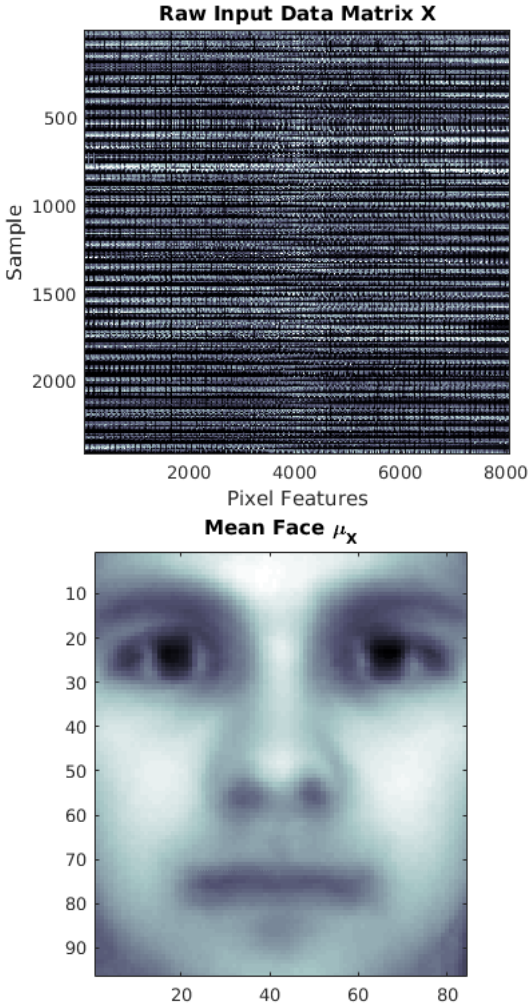

Figure 2. Input Data Matrix and the Mean Face

parameter λ_r is set to ∞. As a consequence, the *soft thresholding operator* is applied to estimating β_k, instead of elastic net. This approach is more accurate and efficient.

1	$K < p$ is the number of sparse PC-loading vectors to estimate
2	$\mathbf{A}_{p\times K}$ contains th first K ordinary PC-loading vectors
3	**for** $k = 1, \cdots, K$ do
4	**while** sparse loading vector β_k still converging **do**
5	**if** $\lambda_r = \infty$ **then**
6	$\beta_k = \cdots$ Soft Thresholding Solution
7	**else**
8	$\beta_k = \cdots$ Elastic Net Solution **end if**
9	$\beta_k = \beta_k / \sqrt{\beta_k^T \beta_k}$ (Normalize to unit length)
10	$\alpha_k = (\mathbf{I} - \mathbf{A}_{k-1}\mathbf{A}_{k-1}^T)\mathbf{X}\mathbf{X}^T\beta_k$ (Update $\mathbf{A}_k$)
11	$\alpha_k = \alpha_k / \sqrt{\alpha_k^T \alpha_k}$ (Normalize to unit length) **end while**
12	**end for**
13	Output the Sparse PC loadings $B = [\beta_1, \cdots, \beta_K]$

Table 1. General Sparse PCA Algorithm (Zou et al, 2006)

The elastic net problem to solved in line 8 is

$$\beta_k = \min_{\beta} \|\mathbf{X}\alpha_k - \mathbf{X}\beta\|^2 + \lambda_r\|\beta\|^2 + \lambda\|\beta\|_1. \quad (12)$$

However, for the $p >> n$ samples in this paper, only the soft thresholding solution in line 6 is used,

$$\beta_k = (|\mathbf{X}^T\mathbf{X}\alpha_k| - \lambda)_+ sign(\mathbf{X}^T\mathbf{X}\alpha_k). \quad (13)$$

3. Results and Discussion

This section describes the results of the experimental survey of SPCA using the 'Extended Yale Face Database B' and the 'SpaSM: Toolbox for Sparse Statistical Modeling' The objective was to investigate how increasing levels of sparseness influences the PC-loading structure and PC-score distributions. The input matrix $\mathbf{X}$ was normalized such that the column means were 0 and the column Euclidean lengths were 1. The additional input arguments passed to the SpaSM function *spca_zouhastie()* are shown in Table 2.

Desired # components $\leq p$	3
Ridge parameter λ_r	∞
Desired # non-zero loadings	$5000, 2500, 1000$

Table 2. SpaSM SPCA Function Arguments

The function returns the sparse PC-loadings and variances of SPCA, as well as those of ordinary PCA. Note that ordinary PCA produces 8064 non-zero loadings.

3.1 Sparse Eigenfaces

Figure 3 shows all of the PC-loadings computed during the study. The first, second, and third row correspond to the first, second, and third PC, respectively. The first, second, third and fourth column correspond to $8064, 5000, 2500, 1000$ non-zero loadings, respectively. Hence, sparseness increases to the right, and principality increases to the top.

The ordinary PC-loadings in the first column exhibit many values close to zero, suggesting the potential for sparseness. As the sparseness increases, the magnitude of the values seem to grow around the regions of active PC-loadings. As explained by Zou et al. 2006, SPCA improves the interpretation of which PC-loadings matter and which do not. For all levels of sparseness, the first two PCs always exhibit *horizontal* asymmetry across the face. As the sparseness increases, the assymetry of the remaining active PC-loadings only strengthens. The meaning of the third PC is more difficult to interpret, but does suggest *vertical* asymetry across the face.

3.2 Sparse PC-Scores

Figures 4 and 5 show all of the samples projected onto the (sparse) PC-loadings 1-3 shown in Figure 3. The four scatterplots in Figure 4 display PC1 versus PC2, and are coloured according to light source azimuth (which ranged from -130 to +130 deg). All four scatterplots indicate that the polar angle

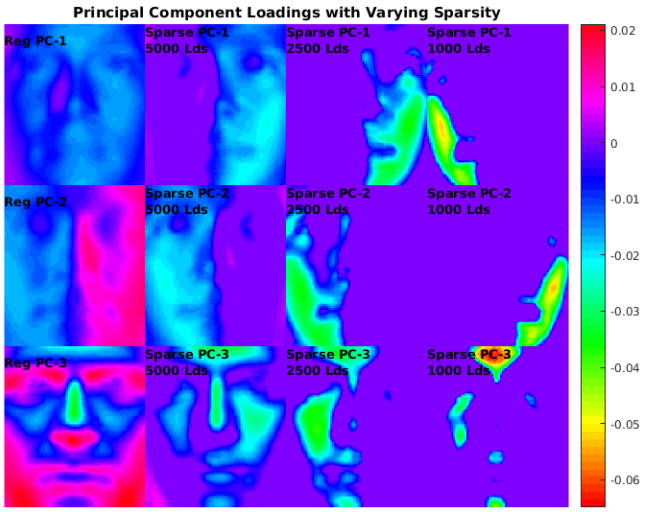

Figure 3. PC-Loadings 1-3 with Varying Levels of Sparseness

(i.e. $tan^{-1}(\frac{PC2}{PC1})$) of each PC-score corresponds to lighting azimuth. It is still unclear what the range (i.e. $\sqrt{PC1^2 + PC2^2}$) of each PC-score indicates.

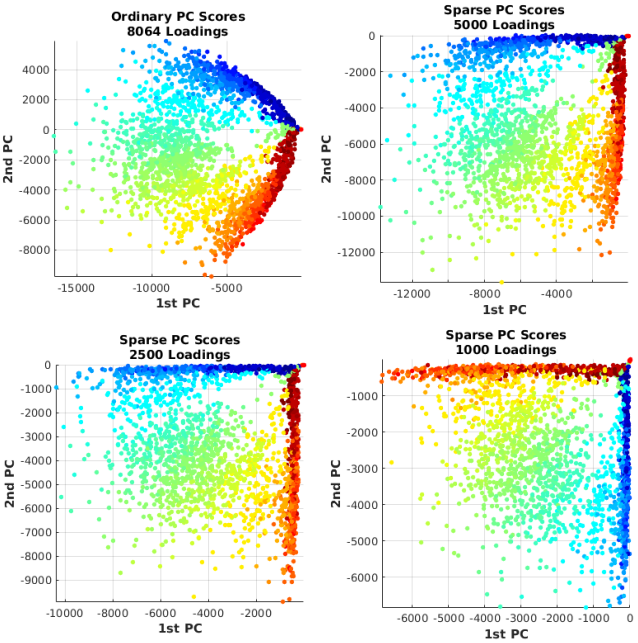

Figure 4. Ordinary and Sparse PC1 vs PC2 Colored by Azimuth $\in [-180 \cdots +180]$ deg

The four scatterplots in Figure 5 display PC1 versus PC3, and are coloured according to light source elevation (which ranged from -45 to +90 deg). Although the sparse PC-scores spread out to greater distances, only the ordinary PC-scores exhibited any seperability in terms of light source elevation. It is still unclear that characteristics of the images these sparse PC-scores are seperated in terms of.

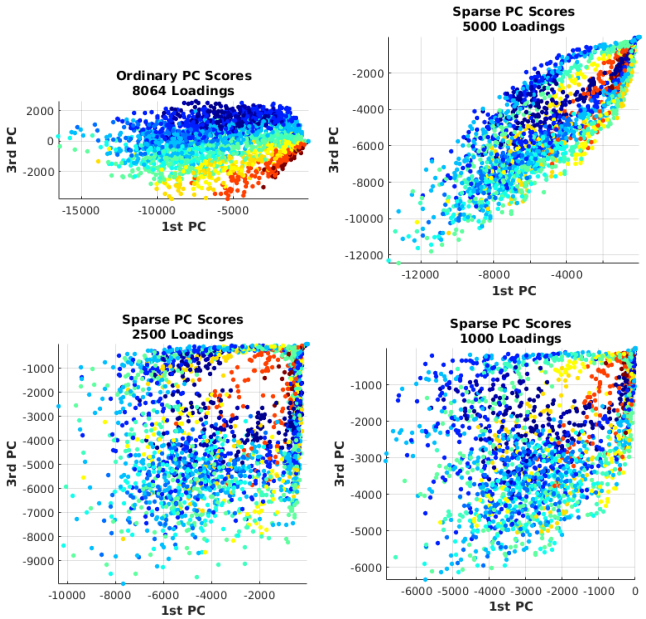

Figure 5. Ordinary and Sparse PC1 vs PC3 Colored by Elevation $\in [-40 \cdots +90]$ deg

4. Conclusion

Ordinary PCA produces 8064 non-zero loading with many values close to zero, which suggests potential for sparsity. As sparsity increases, the magnitude of the values seems to grow around the regions of active PC-loadings. SPCA improves the interpretation of principal components. In this project, for all levels of sparsity, the first two PCs always represent horizontal asymmetry across the face. However, there were still difficulties in interpreting the third PC, which could suggest vertical asymmetry across the face. Relationships among the three PCs and light source were also investigated. Light azimuth at -130 and +130 seem to capture the maximal variances on facial structures, while such phenomena was not observed in change of light source elevation (-45 to +90). These results indicate that sparse regularization of PC-loadings (just as for regression coefficients) can significantly accuracy and interpretability. In this instance, it improved the separation of high-dimensional samples along the axes of their chiefly distinguishing characteristics: light source angles.

5. References

1. Efron B, Hastie T, Johnstone I, Tibshirani R. *Least Angle Regression*. The Annals of Statistics, 32(2), 407–451, 2004.

2. Georghiades, A.S. and Belhumeur, P.N. and Kriegman, D.J. *From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose*. IEEE Trans. Pattern Anal. Mach. Intelligence. vol. 23, no. 6, pp. 643-660, 2001

3. K. Sjöstrand, L.H. Clemmensen, M. Mørup. *SpaSM, a Matlab Toolbox for Sparse Analysis and Modeling*. Journal of Statistical Software, x(x):xxx-xxx, 2010.

4. H. Zou, T. Hastie, and R. Tibshirani. *Sparse Principal Component Analysis*. J. Computational and Graphical Stat. 15(2):265-286, 2006.