SPARSE REPRESENTATIONS FOR HYPERSPECTRAL DATA CLASSIFICATION

Salman Siddiqui, Stefan Robila, Jing Peng, Dajin Wang

Montclair State University
{siddiquis1, robilas, pengj, wangd} @mail.montclair.edu

ABSTRACT

We investigate the use of sparse principal components for representing hyperspectral imagery when performing feature selection. For conventional multispectral data with low dimensionality, dimension reduction can be achieved by using traditional feature selection techniques for producing a subset of features that provide the highest class separability, or by feature extraction techniques via linear transformation. When dealing with hyperspectral data, feature selection is a time consuming task, often requiring exhaustive search of all the feature subset combinations. Instead, feature extraction technique such as PCA is commonly used. Unfortunately, PCA usually involves non-zero linear combinations or `loadings` of all of the data. Sparse principal components are the sets of sparse vectors spanning a low-dimensional space that explain most of the variance present in the data. Our experiments show that sparse principal components having low-dimensionality still characterize the variance in the data. Sparse data representations are generally desirable for hyperspectral images because sparse representations help in human understanding and in classification.

Index Terms— *Hyperspectral data, PCA, SPCA, DSPCA, Sparse representation*

1. INTRODUCTION

Collection and processing of data of all kinds are on scales unimaginable until recently due to exceptional processing power available. Recent advances in high throughput data acquisition, digital storage, computer processing and communication technologies have made it possible to gather, store, and transmit large volumes of data. Hyperspectral imaging is a powerful tool for many real world applications such as agriculture, mining, defense and environmental monitoring. However, hyperspectral imagery tends to be more difficult to process due to high dimensionality [16]. To address this problem, feature extraction techniques such as Principal Component Analysis (PCA) are most often applied. However, in PCA each resulting principal component (PC) is a linear combination of all the original hyperspectral bands. This makes the derived PCs difficult to interpret and the transformed hyperspectral data expensive to classify. To mitigate the problem, rotation techniques and segmented PCA are commonly used. Each has their own drawbacks. Informal thresholding approach used in rotational techniques can be potentially misleading [5], while segmented PCA may not be the most efficient way to segment spectral bands if the goal is to detect a specific target type [6].

There are obvious problems caused by the rapid increase in volume associated with adding extra dimensions to a mathematical space, which is often referred as “Curse of Dimensionality”. The high dimensional data is always difficult to work with for several reasons. Adding more feature means more noise and hence more error. The complexity grows exponentially with the number of dimensions, rapidly outstripping the computational and memory storage capabilities of computers. The curse of dimensionality complicates machine learning problems that involve learning from a finite number of data samples in a high-dimensional feature space.

In most of the cases reducing the number of dimensions improves efficiency. Also, the cost associated with measurement, storage, computation decreases with reduction in dimension. It improves classification performance, helps in interpretation/modeling and also enhances generalization capability. It speed learning process by many folds. Most often, dimension reduction is applied to the high dimensional data. Feature extraction is to apply a map of the multidimensional space into a space of fewer dimensions. This means that the original feature space is transformed by applying e.g. a linear transformation via a principal components analysis. Feature extraction is a method of constructing combinations of the features to get around these problems while still describing the data with sufficient accuracy.

Conventional sparse PCA typically applies simple transforms such as axis rotations and component thresholding [7]. In Zou, Hastie and Tibshirani’s [10] approach called sparse PCA (SPCA), it finds modified components with zero loading in very large problems, by writing PCA as a regression-type optimization problem. This allows the application of Lasso [14], a penalization technique based on the ℓ_1 norm. All these methods are either significantly suboptimal (thresholding) or nonconvex. Direct sparse PCA (DSPCA) improves the sparsity of the principal components by directly incorporating a sparsity criterion in the PCA problem formulation, then forming a convex relaxation of the problem that is a semidefinite program. On the other hand, Segmented PCA [17], a hyperspectral image cube is divided into several non-overlapping blocks in accordance with band-to-band cross-correlations, followed by PCA performed in each block.

2. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA transforms the data into a new co-ordinate system such that the first co-ordinate have maximum variance, second have the second largest and so on. There are two very important optimal properties of PCA [10] which mainly contribute towards its success. First, PCA guarantees minimal information loss by sequentially capturing the maximum variability among variables. Second, PCs are not correlated.

PCA also has an obvious drawback, i.e., each PC is a linear combination of all *n* variables and the loadings are typically nonzero. This makes it often difficult to interpret the derived PCs.

Principal Component Analysis is usually done by Singular Value Decomposition (SVD). In detail, let the data *Z* be an (*m*, *n*) matrix where *m* is number of samples and *n* is number of features. Then we have the SVD of *Z* as

$$Z = UDV^T \tag{1}$$

Where ^{T} stands for transpose, *U* are the principal components (PC's) of unit length and the columns of *V* are the corresponding loadings of the principal components. The eigenvalue decomposition of the covariance matrix can be written as

$$A = Z^T Z = VD^2V^T \tag{2}$$

$$A = \sum_{i=1}^{n} \varepsilon_i V_i V_i^T \tag{3}$$

where V_i are the eigenvector and $\varepsilon_i = D_{i,i}$ are the corresponding eigenvalues. We can also note that PCA is orthogonal transformation of the data.

3. SPARSE PCA

In this paper we experiment with SPCA and DSPCA technique on the hyperspectral and sonar data. We will also analyze the sparse representation of the datasets and its effect on classification. The following are short descriptions of the two techniques.

3.1 SPCA

SPCA is built on the fact that PCA can be written as a regression-type optimization problem. Thus they have integrated Lasso (elastic net) directly into the regression criterion such that the resulting modified PCA produces sparse loadings. In general it does not replace regular PCA but SPCA modifies it to produce sparse principal components. SPCA is a regression optimization framework in which PCA is done exactly, unlike linear regression and least square solution. The modified regression framework allows a direct modification by using the Lasso (elastic net) penalty such that the derived loadings are sparse. The advantage of using Lasso is that it continuously shrinks the coefficients to zero. It produces a sparse model using a variable selection method. The numbers of selected features are limited by the number of samples. It selects only one of the highly correlated features and does not care which one is in the final model. On the other hand, elastic nets [12] incorporate ridge penalties as well as lasso penalty. Limitation of Lasso [14] is removed by adding a ridge constraint with the Lasso constraint so that all variables are included in the model. In the SPCA framework, Equation (1) is extended as follows. For all *i,* denote

$$Y_i = U_i D_{i,i} \tag{4}$$

where Y_i is the i^{th} principal component of *Z*. Then, solve:

$$\hat{\beta} = \underset{\beta}{\arg\min} |Y_i - Z\beta|^2 + \lambda|\beta|^2 + \lambda_1|\beta| \tag{5}$$

where β is a regression coefficient, the second term $\lambda|\beta|^2$ is the ridge penalty and the third term $\lambda_1|\beta|$ in Equation (5) is the lasso penalty. SPCA is a reconstruction of PCA in the regression framework.

3.2 DSPCA

A direct formulation of SPCA analysis (DSPCA) [3] produces dominant principal components and factors ranked according to their variance for a given covariance matrix of the input data. Numerically, the eigenvalue decomposition of the covariance matrix is given by equation (3), for sparsity it can be written as

$$\begin{aligned} \max \quad & x^T Ax \\ \text{subject to} \quad & \|x\|_2 = 1 \\ & \mathrm{Card}(x) \le k \end{aligned} \tag{6}$$

where *Card(x)* denotes the cardinality (number of non-zero elements) of ***x*.** Equation (6) is non-convex and numerically hard to solve in polynomial time. To solve the above equation more efficiently, d' Aspremont et. al. have reformulated the optimization problem into the following optimization problem.

$$\begin{aligned} \max \quad & \operatorname{Tr}(AX) \qquad (7) \\ \text{subject to} \quad & \operatorname{Tr}(X) = 1 \\ & 1^T |X| 1 \le k \\ & X \succeq 0 \end{aligned}$$

where $X = xx^T$.

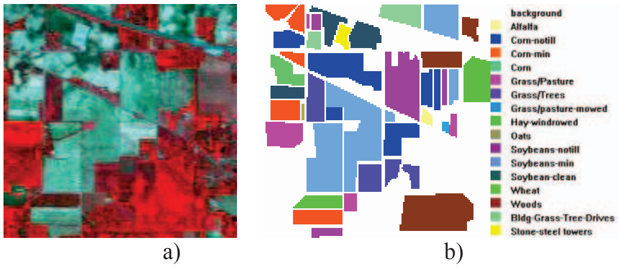
Note that $Card(x) \le k$ is replaced by the weaker but convex constraint i.e. $1^T |X| 1 \le k$ (ℓ_1 norm of x), where $|X|$ takes the absolute values of X. The rank constraint was also dropped. For small problems, DSPCA uses the interior point method. But for large problems with high dimensionality, a first order minimization technique is applied to the semidefinite program arising in the semidefinite relaxation of sparse PCA.

4. EXPERIMENTAL RESULTS

In order to understand how sparse representation affects feature extraction, we have applied both algorithms (SPCA and DSPCA) to two datasets. The first (Figure 1a) is an AVIRIS hyperspectral image [11] of 220 bands (from 400 to 2500nm) and 145 by 145 pixels. The image is accompanied by ground truth that identifies 17 classes corresponding to various crops, roads, and man-made structures (see Figure 1b). The second is a sonar dataset used by Gorman and Sejnowski [9] in their study of classification of sonar signals using a neural network. The classification accuracy of Sonar data set [1] using all 60 features was 82.7%. We have used only 25% of the total features and sparsity was set to 50%. Figures 2 and 3, show the amplitude of loadings for the PCA, SPCA and DSPCA. Here, the vertical axis corresponds to the magnitude of each principal component produced by PCA, SPCA and DSPCA, and the horizontal axis corresponds to the resulting components. We note that some of the features produced by PCA were selected as it is by SPCA. The reason for that is SPCA uses the same principal components produced by regular PCA and applies Lasso and elastic nets [12] for sparse loading.

Next we used the resulting components to classify the data using k nearest neighbor classification. Figures 4 and 5 show the classification accuracy versus the number of components used to classify. The accuracy is on average the same as sparse PCA in which only 50 % non-zero loadings are present. We note that the accuracy with only half of the features used for classification is approximately the same as with full features. Due to the computational constraint of DSPCA we are using only a quarter of the hyperspectral features. During analysis with PCA, we cannot determine which features are contributing, but with Figure 2 and Figure 3, SPCA and DSPCA shows which features are actually contributing.

The results suggest that the accuracy values for SPCA and DSPCA are at least equal to the regular PCA. Also Figures 2 and 3 clearly indicate which features actually contribute in classification. We can also note that with each additional feature the accuracy improves but after certain number the accuracy does not change or even deteriorates.



a) b)

Figure 1: AVIRIS data scene a) color composite image, b) ground data for the scene and description of the classes

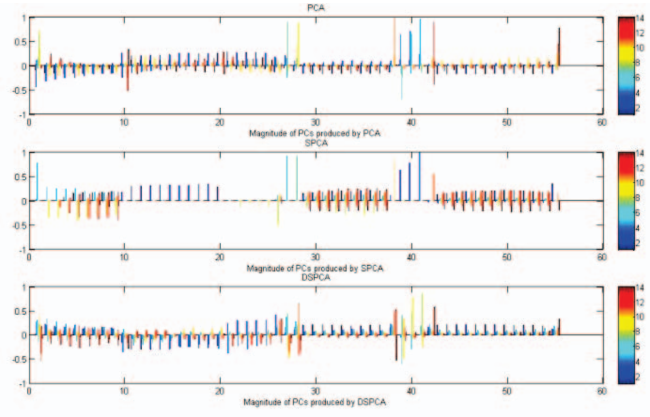


Figure 2: Magnitude of Principal Components used for classification on Hyperspectral data

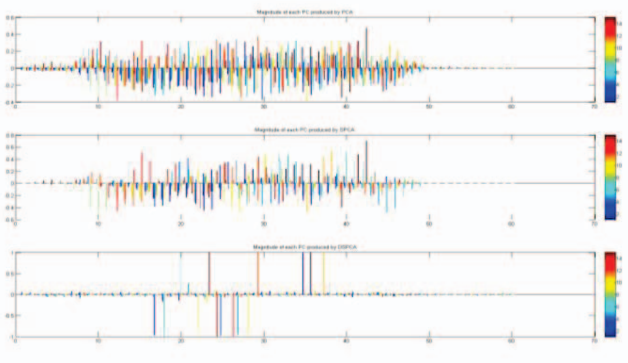


Figure 3: Magnitude of Principal Components used for classification on Sonar data

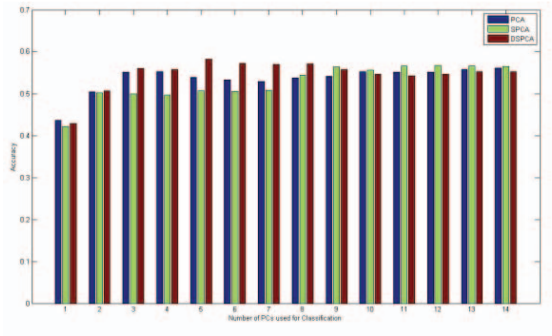


Figure 4: Comparison of Accuracy obtained with PCA, SPCA and DSPCA on Hyperspectral data

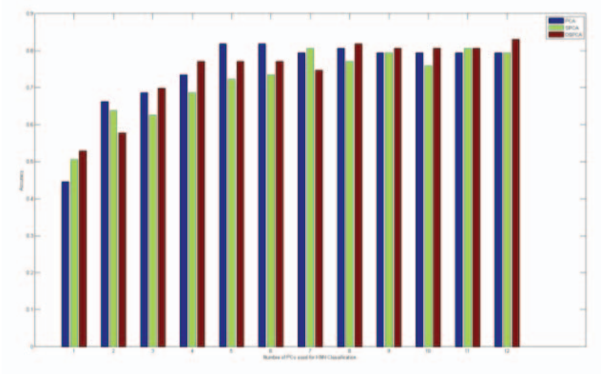


Figure 4: Comparison of Accuracy obtained with PCA, SPCA and DSPCA on Sonar data

Compared to PCA SPCA and DSPCA require a lot of processing, but where the transaction cost is high and analyzing datasets is of primary importance. It applies the Lasso and elastic net technique to obtain sparse loading from the regular principal components from PCA. Without any sparsity constraint, SPCA reduces to the regular PCA. It also avoids misidentification of important variables. On the other hand DSPCA with Sedumi [13], the interior point method is less computationally efficient and cannot be used for high dimensional datasets.

5. CONCLUSIONS

After experimenting with Hyperspectral and Sonar data, we can conclude that sparse data representations are generally desirable because a sparse representation helps in human understanding and helps in classification. For very high dimensional data SPCA outperforms DSPCA. It is also computationally efficient for both small and large dataset.

6. REFERENCES

[1] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Bharath K. Sriperumbudur, David A. Torres, Gert R. G. Lanckriet, "Sparse eigen methods by D.C. programming", *ACM International Conference Proceeding Series; Vol. 227*

[3]d'Aspremont, L. El Ghaoui, M. I. Jordan, G. R. G. Lanckriet "A Direct Formulation for Sparse PCA using Semidefinite Programming", *SIAM Review*, 49(3), pp. 434-448, July 2007.

[4] d'Aspremont, F. Bach, L. El Ghaoui, "Optimal Solutions for Sparse Principal Component Analysis." ICML proceedings

[5] Du, Qian; Chang, Chein-I, "Segmented PCA-based compression for hyperspectral image analysis", *Proceedings of the SPIE*, Volume 5268, pp. 274-281

[6] F. Tsai, E.-k. Lin and K. Yoshino, Spectrally segmented principal component analysis of hyperspectral imagery for mapping invasive plant species, *International Journal of Remote Sensing,* 2006

[7] J. Cadima and I. T. Jolliffe, Loadings and correlations in the interpretation of principal components, *Journal of Applied Statistics*, 22 (1995), pp. 203–214.

[8] Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12, 531–547.

[9] Gorman, R. P., and Sejnowski, T. J. (1988). "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets" in Neural Networks, Vol. 1, pp. 75-89.

[10] Hui Zou, Trevor Hastie, Robert Tibshirani, "Sparse Principal Component Analysis" (Technical Report). *April 26, 2004*

[11] Landgrebe, Larry Biehl, Purdue University, West Lafayette, IN [http://cobweb.ecn.purdue.edu/], 220 band dataset.

[12] LARS, the Elastic Net and SPCA (Technical Report). Informatics and Mathematical Modelling, Technical University of Denmark.R.

[13] Sturm J, Using SEDUMI 1.0x, a MATLAB toolbox for optimization over symmetric cones, Optimization Methods and Software, 11 (1999), pp. 625–653.

[14] Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal statistical society, series B, 58 (1996), pp. 267–288.

[15] Vandenberghe, L., & Boyd, S. Semidefinite programming(1996).SIAM Review, 49–95.

[16] Vassilis Tsagaris, Vassilis Anastassopoulos, and George A. Lampropoulos, "Fusion of Hyperspectral Data Using Segmented PCT for Color Representation and Classification", *IEEE Transactions on GeoScience and Remote Sensing, Vol. 43, No. 10, October 2005*

[17] Xiuping Jia, John A. Richards, "Segmented Principal Components Transformation for Efficient Hyperspectral Remote-Sensing Image Display and Classification", *IEEE Transactions on GeoScience and Remote Sensing, Vol. 37, No.1, January 1999*