CS 736 Project
Medical Image Processing

Project Topic

PLS Dimension Reduction
Applications in Classification

Team

Aditya Kumar Akash, 120050046
Praveen Agrawal, 12D020030

Abstract

In this project we understand the Partial Least Squares method for dimensionality reduction. We then apply the PLS regression onto different dataset as a classification tool. The comparison of PLS regression to other techniques is also studied.
We also see PLS as a tool for data visualization.

Introduction

PLS (Partial Least Squares) is a dimension reduction technique. PLS is used to find the fundamental relations between two matrices (*X* and *Y*), i.e. a latent variable approach to modeling the covariance structures in these two spaces.
It reduces the dimensionality of a data set by projecting the data onto components of maximum covariance with a second data set.

The goal of PLS regression is to predict Y from X and to describe their common structure. When Y is a vector and X is a full rank matrix, this goal could be accomplished using ordinary multiple regression. When the number of predictors is large compared to the number of observations, X is likely to be singular and the regression approach would not be feasible. It is characteristic of domains such as, e.g., bioinformatics, brain imaging and genomics.

PLS assumes the following fundamental relationship between the two sets :

$$X = TP^T + E$$
$$Y = UQ^T + F$$

where X is an $n \times m$ matrix of predictors, Y is an $n \times p$ matrix of responses; T and U are $n \times l$ matrices that are, respectively, projections of X (the *X score*, *component* or *factor* matrix) and projections of Y (the *Y scores*); P and Q are, respectively, $m \times l$ and $p \times l$ orthogonal *loading* matrices; and matrices E and F are the error terms, assumed to be independent and identically distributed random normal variables. In PLS the covariance between score vectors is maximized in each iteration.

PLS1 Algorithm

PLS1 is a widely used algorithm appropriate for the vector *Y* case. It estimates *T* as an orthonormal matrix. Following is the pseudo code :

```
**function** $PLS1(X, y, l)$
    // $X$ is the dataset in consideration
    **for** $k = 1$ **to** $l$
        // initial weight which maximizes covariance
        $w^{(k)} \leftarrow X^T y$
        $w^{(k)} w^{(k)} / ||w^{(k)}||$

        // score estimation
        $t^{(k)} \leftarrow Xw^{(k)}$
        $c_k \leftarrow {t^{(k)}}^T t^{(k)}$
        $t^{(k)} \leftarrow t^{(k)} / c_k$

        // estimating the loading weights based on scores
        $p^{(k)} \leftarrow X^T t^{(k)}$
        $q_k \leftarrow y^T t^{(k)}$
        If $q == 0$
            $k \leftarrow i$
            break the loop

        // deflating $X$
        $X \leftarrow X - c_k t^{(k)} p^{(k)^T}$
    end for
define W to be the matrix with columns $w^{(0)}, w^{(1)}, ..., w^{(l-1)}$.
Do the same to form the P matrix and q vector.
$B \leftarrow W(P^T W)^{-1} q$
$B_0 \leftarrow q_0 - P^{(0)^T} B$
return $B, B_0$
end function.
```

Experiments

We used PLS regression as a classification tool. We used MNIST dataset, and two microarray dataset to demonstrate the utility of PLS. Following are the description of experiments.

MNIST Dataset

The classification of handwritten digits using PLS is performed to test the working when number of instances are sufficient.

The dataset consists of 28x28 images. The vectorized image is taken as the feature vector. 60,000 instances are taken for training such that approximately 6,000 instances are present for each digit. The test data contains 10,000 images.

PLS was used as one vs all classifier. 10 classifiers were built for each digit and we combined the predictions of each of them to get the prediction.
Following are the accuracy of classification using PLS using different number of

Number of Components	Accuracy
3	82.340000
5	84.680000
10	85.500000

Further increasing the number of components does not increase the accuracy by large amounts.

It is worth noting that PLS achieves good accuracy at 4-5 components. While a PCA with factor analysis based implementation using Mahalanobis distance for classification uses 8-9 latent dimension to get such accuracy. Thus aligning the components to increase covariance is efficient way to reduce dimension.

Microarray Dataset

High-dimensionality makes the application of most classification methods difficult. The presence of very few number of samples compared to features for each sample poses unnecessary complication for most classification methods. Such scenarios often arise in gene related experiments where a lot of genes are involved in decision of a particular trait / disease.
Influenced from “ PLS Dimension Reduction for Classification with Microarray Data ” paper, we compare the accuracy of PLS with an SVM classifier.

We use following datasets

1. Prostate dataset : Singh (2002), 102 samples, 2135 gene features
   50 normal, 52 tumor samples
2. DLBC Lymphoma Dataset : 77 samples, 2645 gene features
   19 normal, 58 positive samples

The SVM classifier used is the standard svmtrain classifier provided by MATLAB.

For each dataset we conduct 100 experiments to find the average accuracy. The dataset is divided randomly into 7 : 3 ratio for training and testing. Number of components is found by improving the training classification.

Following are the results

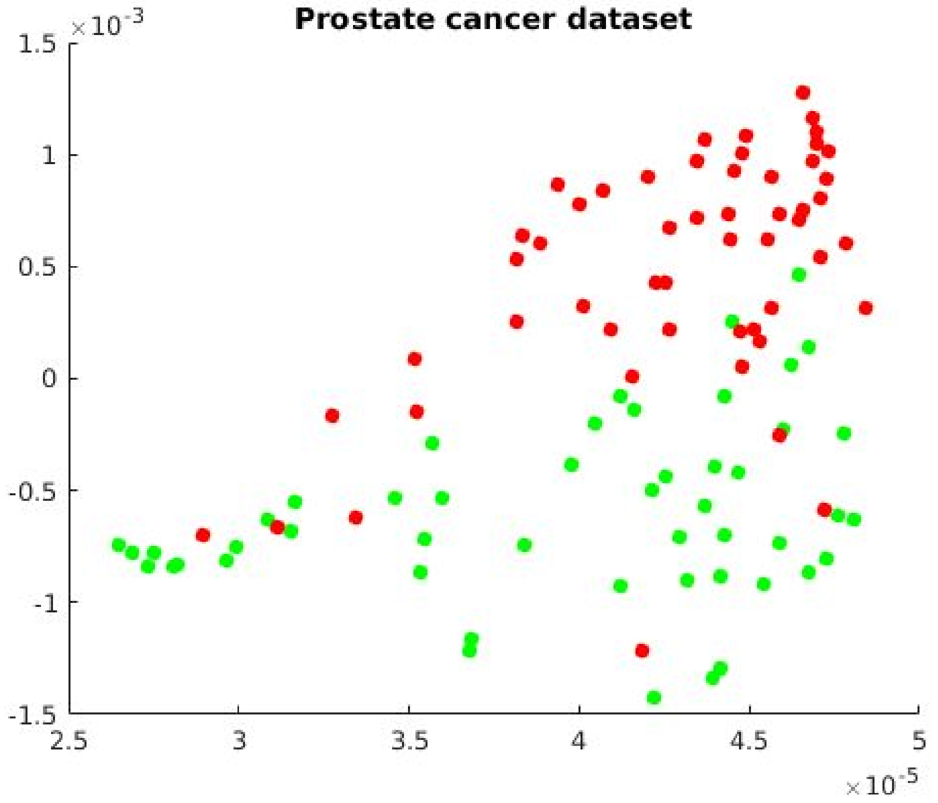
Dataset	Optimal Component	SVM Accuracy	PLS Regression Accuracy
Prostate	5	88.893530	91.081492
DLBCL	3	94.901315	96.196989

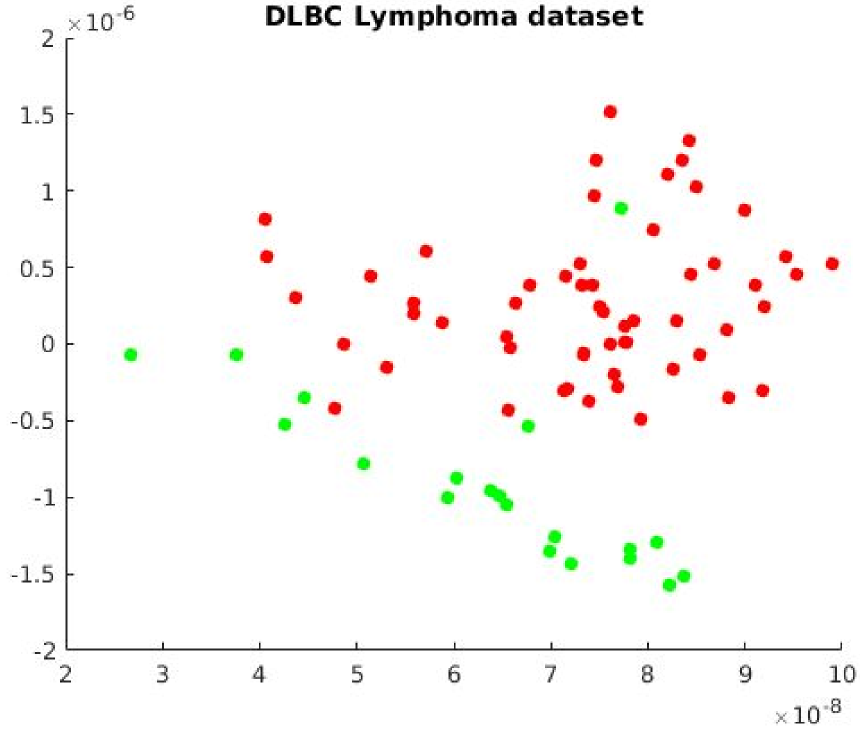
For these dataset we can see that PLS regression classifier performs slightly better than a standard SVM classifier.
The paper also points to PLS being as good a classifier for these kind of dataset as any other classifier. This provided motivation for understanding PLS and using it for other application as feature selection.

Data Visualization

PLS can also used for data visualization. This gives insight into how the data is correlated to the observations and could help in selecting better features / designing better classifier.

We made scatter plot for the complete datasets.
Red represents the sample points which have the disease, while green represents the other case. Clearly there is a separation between the two cases. So using a SVM with good features / PLS scores could lead to even better classification.


Prostate cancer dataset
×10^-3
1.5
1
0.5
0
-0.5
-1
-1.5
2.5
3
3.5
4
4.5
5
×10^-5


DLBC Lymphoma dataset
×10^-6
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
2
3
4
5
6
7
8
9
10
×10^-8

Conclusion

PLS proves to be a powerful tool when it comes to classification of data where we have very less number of instances compared to the features for training. PLS is also fast in training as well as prediction as it involved linear operations. It also shows how considering the correlating in data helps in better understanding (here prediction).

We learnt a great deal in completing the project. There is still need for good classifiers which would help the medical world in fast and accurate diagnosis of chronic diseases.

We would like to thank Prof. Suyash for giving us the opportunity to undertake this study.

References

https://en.wikipedia.org/wiki/Partial_least_squares_regression

PLS Dimension Reduction for Classification with Microarray Data, Anne-Laure Boulesteix, http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/pdf/sagmb.pdf

https://en.wikipedia.org/wiki/Microarray_databases

http://yann.lecun.com/exdb/mnist/