Introduction

Dataset prostate gene expression data from micro-array experiment (p= 6033 genes, n=102 subjects)

Goal Predict whether a person has prostate cancer or not based on gene expression data

Methods Used

Following approaches were used for the analysis:

- Dimensionality reduction
  - PCA
  - Supervised PCA
- Classification
  - LDA
  - QDA

Dimensionality reduction

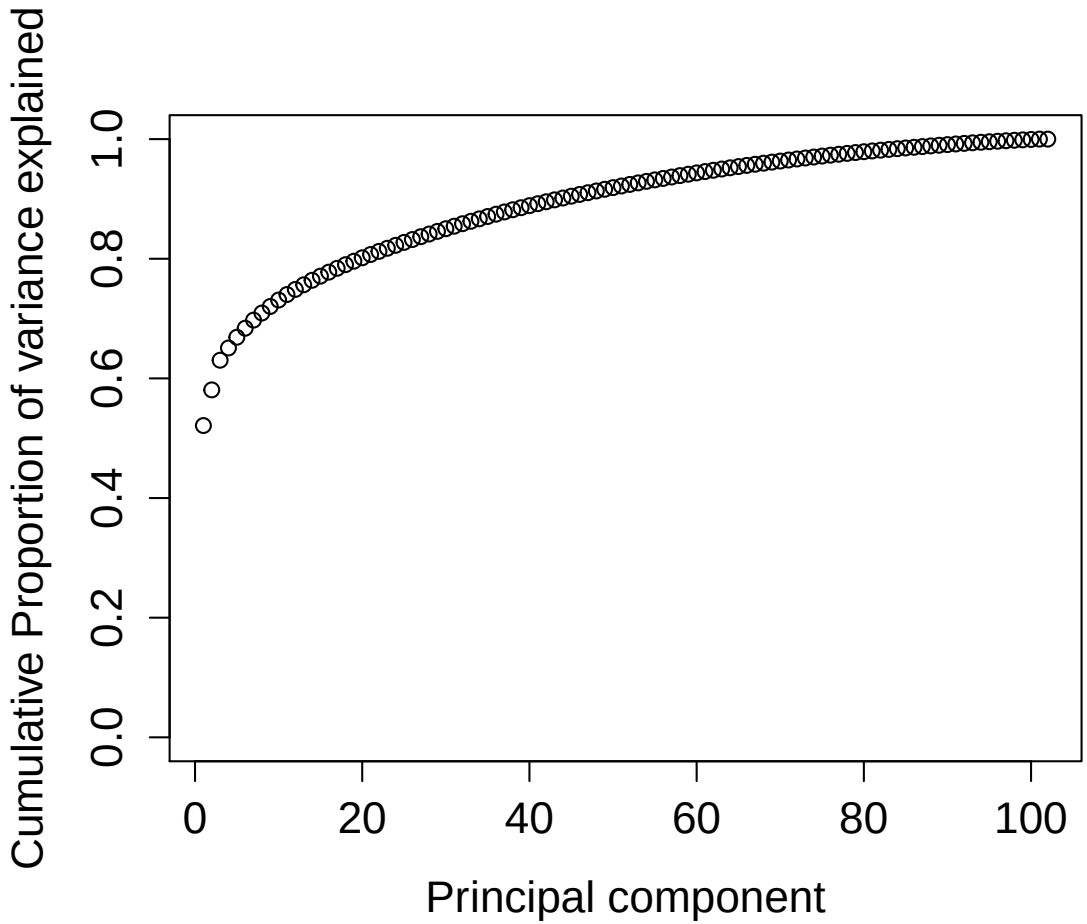

Figure: PCA Analysis

Dimensionality reduction and Classification

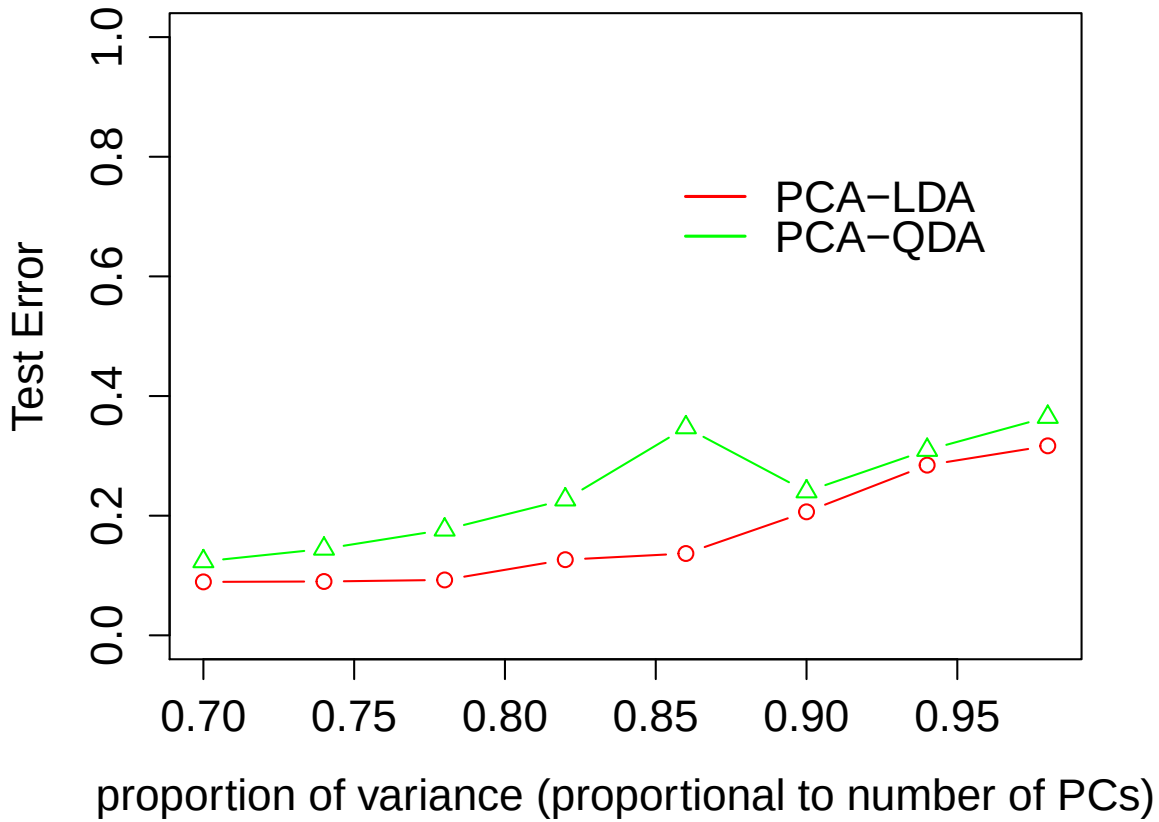

Figure: PCA with classification methods

Supervised PCA for Dimensionality reduction

Steps for Supervised PCA:

- Compute (univariate) standard regression coefficients for each feature.
- Form a reduced data matrix consisting of only those features whose univariate coefficient exceeds a threshold θ in absolute value.
- Compute the first (or first few) principal components of the reduced data matrix.
- Use these principal component(s) in a regression model to predict the outcome

Supervised PCA and Classification

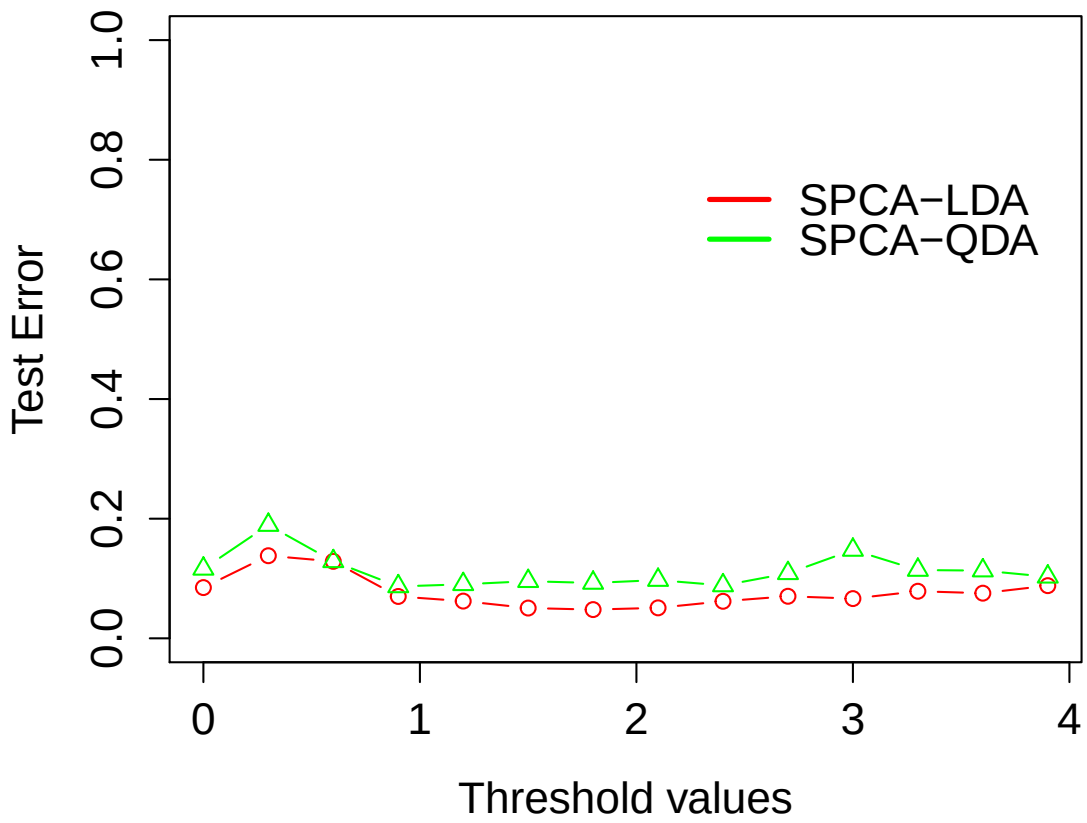

Figure: SPCA and classification using 70% of variability

Supervised PCA and Classification

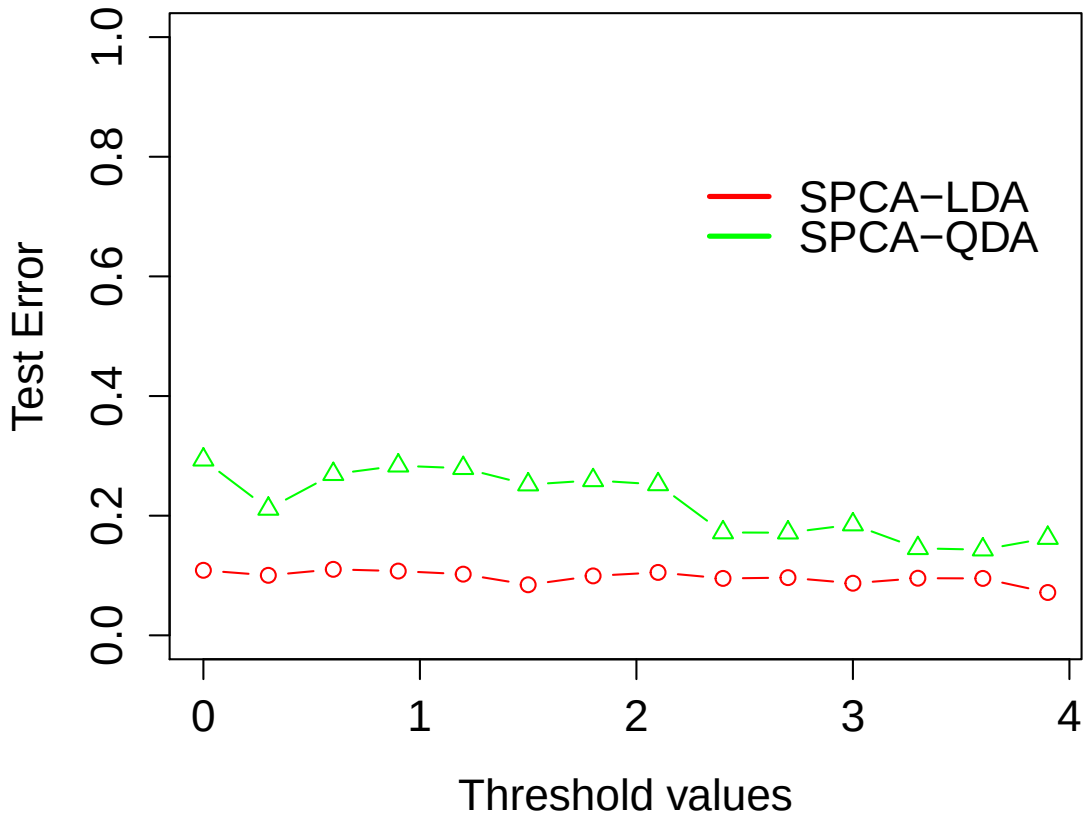

Figure: SPCA and classification using 85% of variability

Supervised PCA and Classification

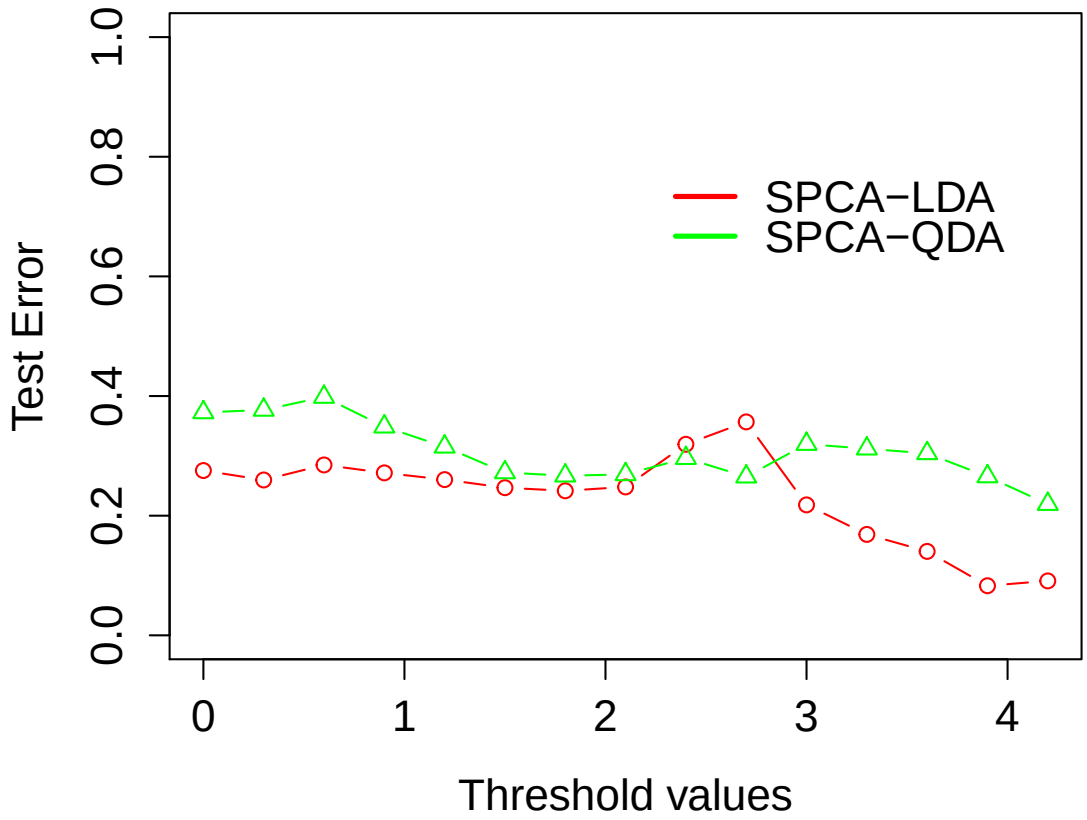

Figure: SPCA and classification using 95% of variability

Supervised PCA - Threshold Selection

Variability	PCs	Threshold	No. of genes	Test Error
70%	11	1.5	601	0.054
95%	14	3.9	25	0.082

Table: SPCA Threshold values

Supervised PCA and Classification

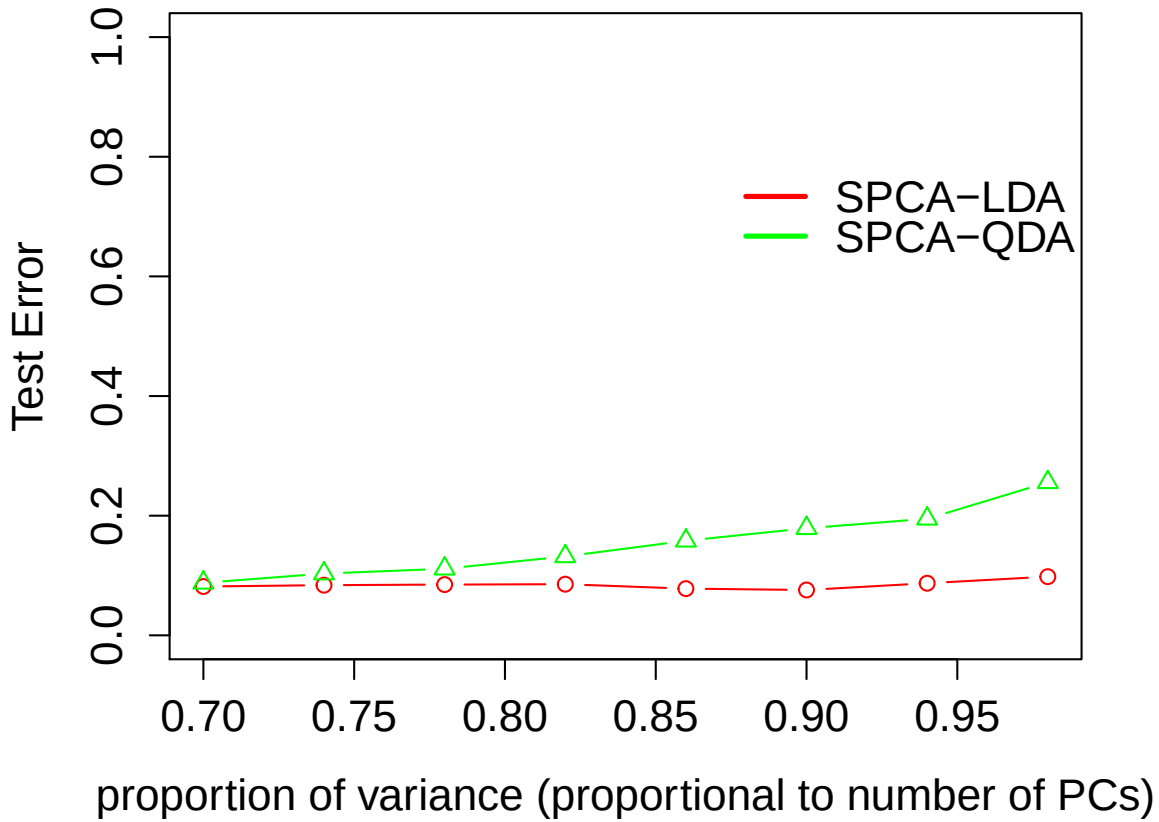

Figure: SPCA and classification using threshold value of 3.9

Conclusions - Test Error

Method	No of genes	PCs	Test Error
PCA-LDA	6033	7	0.088
SPCA-LDA	25	10	0.072

Table: SPCA Threshold values