Application of Various Classification Methods for Prostate Cancer Prediction Using Microarray Data

PCA, sPCA, LDA, SPLS-DA, NSCC

Tianyao Lu, Garima Garg

AMS 225

12/01/2015

Introduction & Background

Microarray Uses mRNA/cDNA onto a glass slide for measuring expressivity (signal).

Dataset This microarray data is prostate cancer specific which consists of 102 participants and 6033 different genes. The signal was processed using log 10-base transformation and then standardized.

Train - Test Data Train and test data were not available, so we took random sample 70/32 for training and testing, respectively.

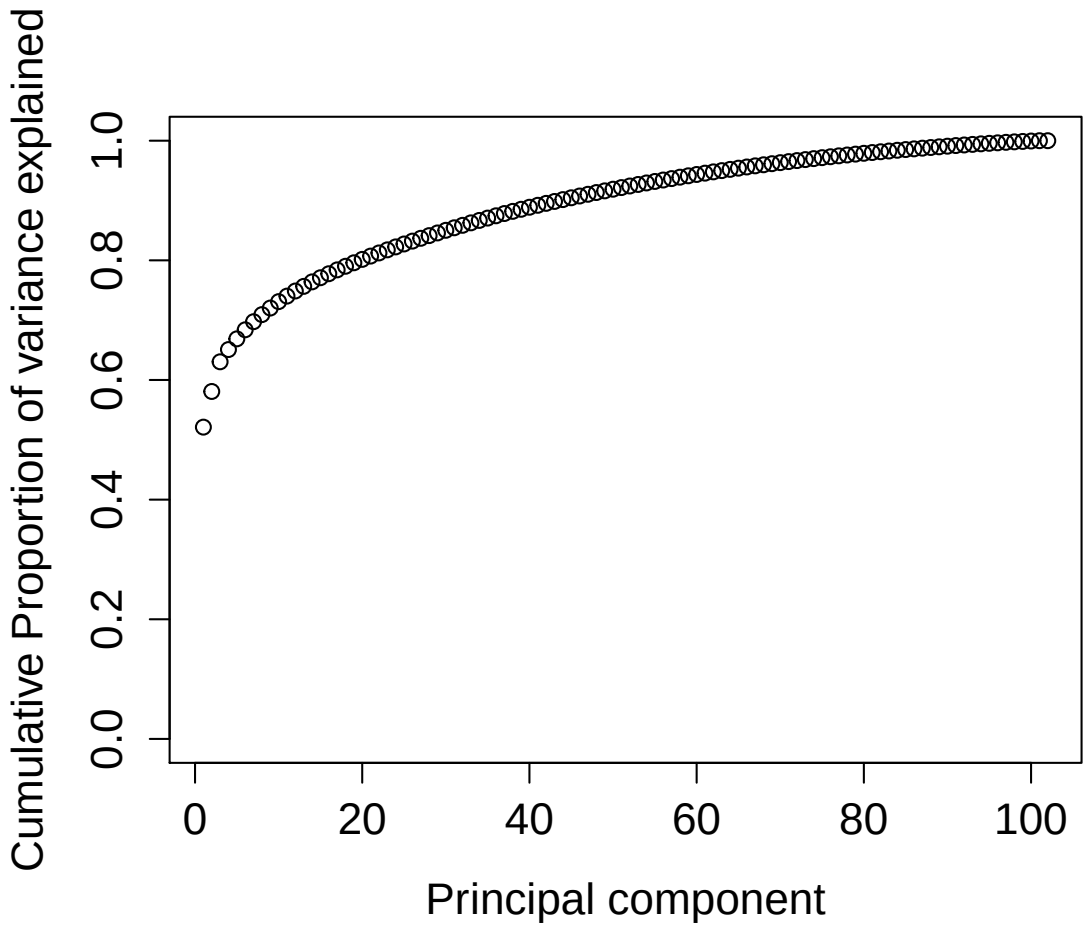
Goal Find a classifier that works well* to predict prostate cancer status (maybe useful as a clinical guideline).

Methods

- Dimension Reduction Methods: PCA, Supervised PCA
- Classification Methods: LDA, QDA*
- Classification with build in dimension reduction feature: SPLS-DA*, NSCC

Dimensionality reduction - PCA Analyis

Figure: Number of PC scores = min(n,p) = 102



Dimensionality reduction and Classification

Steps to be followed:

- Sample 70% of data(training data), to build LDA and QDA classifiers using PC scores (How many PC scores to include?).
- Use the classifiers to make predictions on the holdout sample(test data).
- Compute test error.
- Repeat this process 100 times to get the averaged error rate.

Dimensionality reduction and Classification

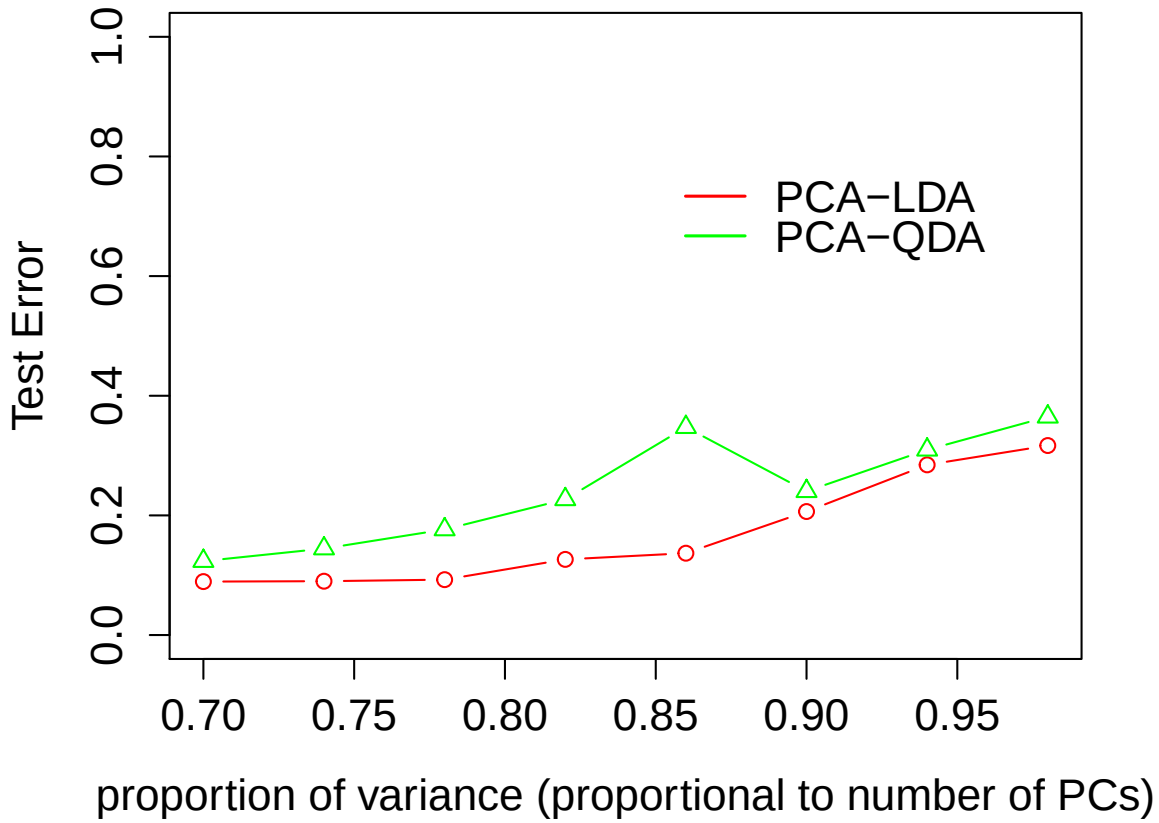


Figure: PCA with classification methods

Supervised PCA - Identify key genes

Steps for Supervised PCA:

- Compute (univariate) standard regression coefficients for each feature.(In our case, use logistic regression).
- 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 - How to choose threshold value?

Steps to be followed:

- For each of the threshold value,
  - Get the reduced data matrix.
  - Compute PC scores on reduced data matrix.(How many to include?).
  - Sample 70% of data to build LDA and QDA classifiers, and use them for predictions on holdout samples to compute test error. Repeat the process 100 times to get the averaged error rate.
- Choose the threshold value that gives you minimum test error rate.

Supervised PCA - Threshold value

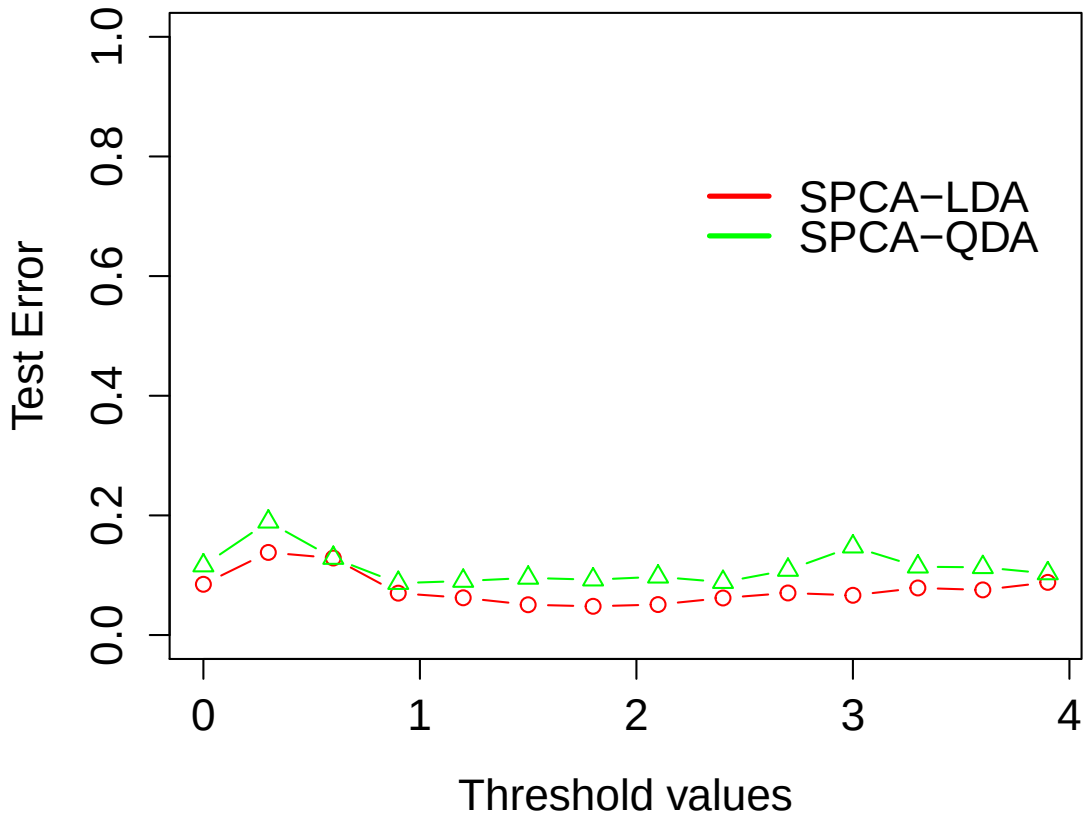


Figure: SPCA and classification using 70% of variability

Supervised PCA - Threshold value

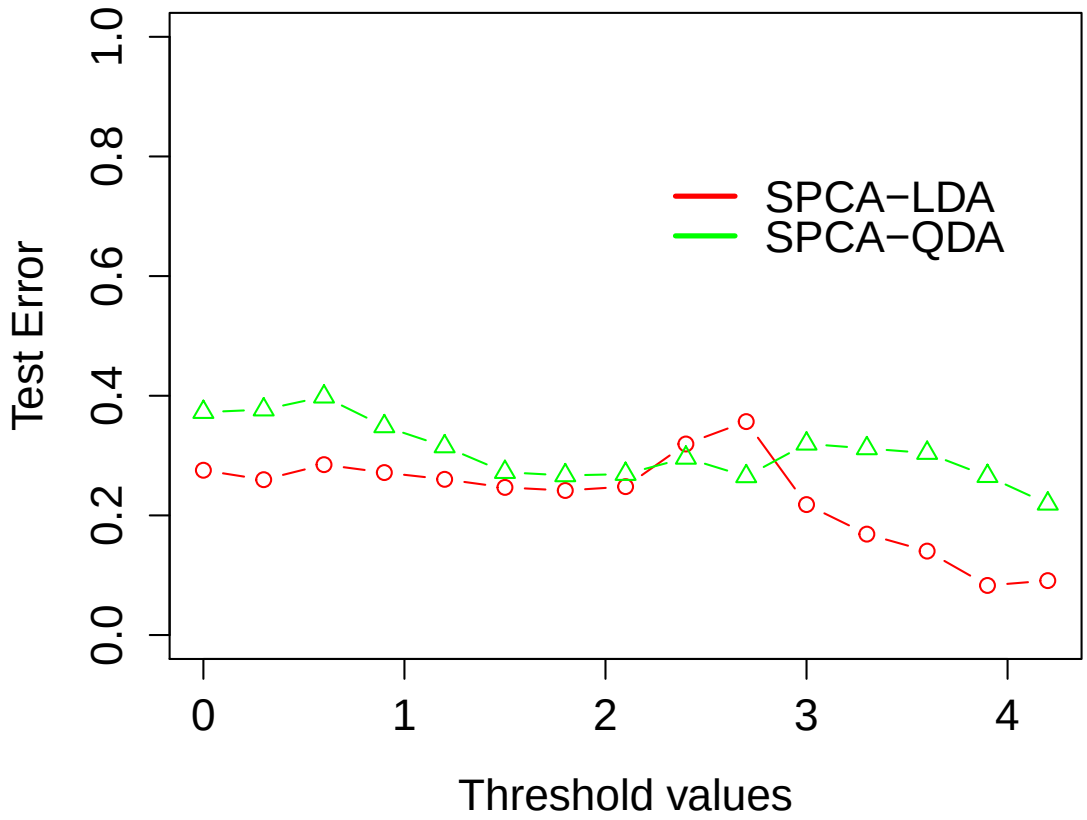


Figure: SPCA and classification using 95% of variability

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 - Choosing Threshold value

Supervised PCA and Classification

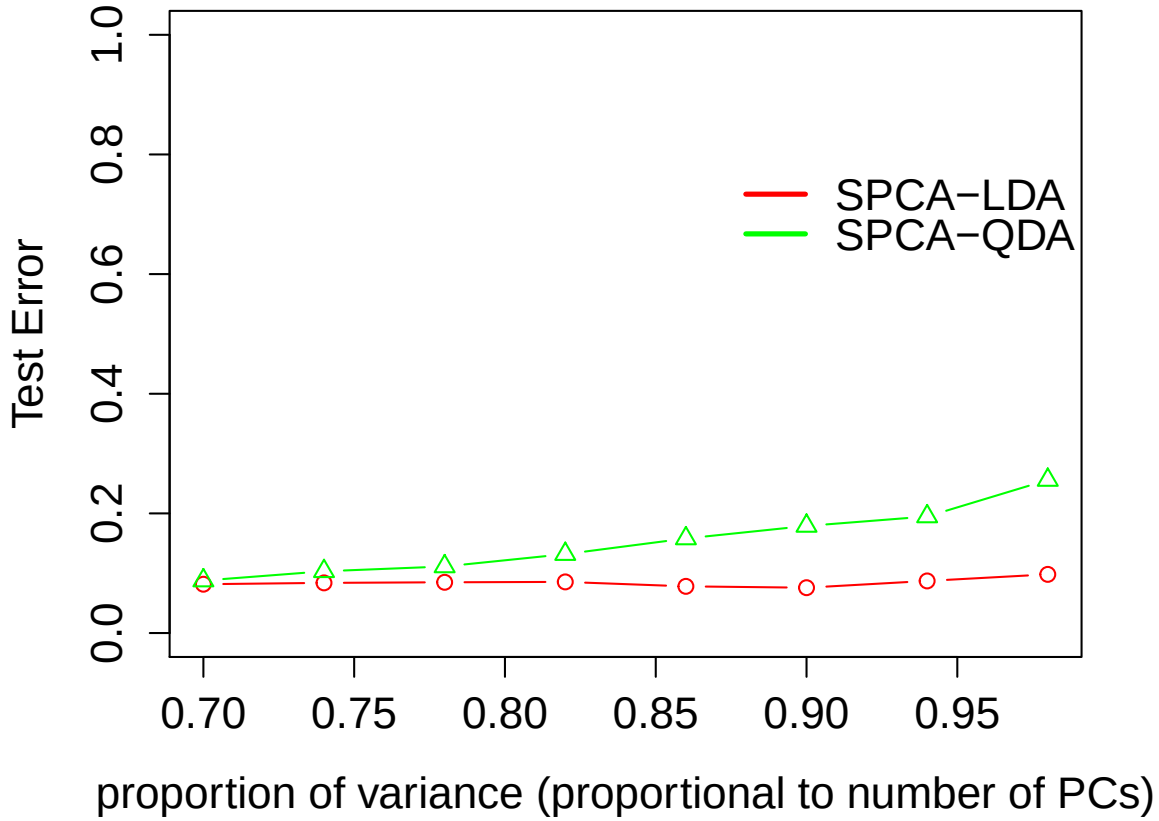


Figure: SPCA and classification using threshold value of 3.9

Conclusions - Test Error

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

Table: PCA and SPCA test results

LDA for prostate microarray data

1. LDA is simple, fast, and reasonably reliable for standard datasets.
2. p $>>$ n dataset will challenge LDA in both prediction and estimation.
3. Multicollinearity from genes will affect the performance of LDA (R warned me every time it runs).
4. It produces inconsistent parameter estimations due to lack of sample size.
5. It makes matrix operation harder.
6. LDA prediction is not reliable using our dataset.

LDA Results

- Prevalence is abour 14% (seer.cancer.gov).
- Set prior probability as 14%.

Build LDA classifier on training data and we make prediction using testing data.

Table: LDA Confusion Matrix

	Truth	
Pred	0	1
0	15	10
1	0	7

Based on this training data, our accuracy is only about 0.69 (Noted random guess is 50%).

Introducing Sparse PLS-DA

What is PLS and PLS-DA

1. It is based on partial least squares which is one method for multivariate regression when p >> n.
2. PLS developed into classification by establishing connecting with Fisher's LDA (PLSDA).
    a. Treat response as continuous and PLS will develop K latent variables that are linear combinations of originals.
    b. Use an off-the-shelf method to perform classification on latent variables since in general $K << n << p$.

Introducing Sparse PLS-DA Cont.

What is PLS and PLS-DA Cont.

Suppose we have $\mathbf{Y}_{n\times q}$ and $\boldsymbol{X}_{n\times p}$, PLS assumes underlying latent components $\mathbf{T}_{n\times K}$. We have the following model,

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F}$$

$$\boldsymbol{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

Where $\mathbf{E}$ and $\mathbf{F}$ are errors, $\mathbf{P}$, $\mathbf{Q}$ are loadings, and $\mathbf{T}$ is latent component. Simply, we have $\mathbf{T} = \boldsymbol{X}\boldsymbol{W}$ and $\boldsymbol{W}_{p\times K}$ is matrix consists of K direction vectors $(1 \leq K \leq \min(n, p))$.
Obtain k^{th} direction vector $\boldsymbol{w}_k$ in univariate case by
$\max\{Corr^2(\mathbf{Y}, \boldsymbol{X}\boldsymbol{w})Var(\boldsymbol{X}\boldsymbol{w})\}$ subject to $\boldsymbol{w}^T\boldsymbol{w} = 1$ and $\boldsymbol{w}^T\Sigma_X\hat{w}_l = 0$

Introducing Sparse PLS-DA Cont.

SPLS and SPLS-Classification

It is based previous PLS method. In addition, it incorporates variable selection into PLS by solving the objective function contains L1 and L2 penalties.

1. SPLSDA is a direct application of SPLS and algorithm is similar to PLS-DA (previously discussed).
2. It reduce dimension as well as making prediction which produce better results compared with PLS-DA.
3. Its theoretical derivation shows that we only need to find the very first direction vector for classification – computationally efficient.

Building Classifier using SPLS-DA and PLS-DA

SPLSDA–Finding optimal thresholds

Solution of objective function for SPLS-DA has two thresholds $\eta \in (0, 1)$ and K. They are controlling the amount of shrinkage and total number of direction vectors.

1: Use 90 pairs of (η, K) to build classifier and select the pairs (25 were selected) such that most of genes are eliminated (gene number ≤ 250).

2: Use the rest of pairs to perform prediction on training data and select the best one ($\eta = 0.8, K = 3$).

SPLSDA results

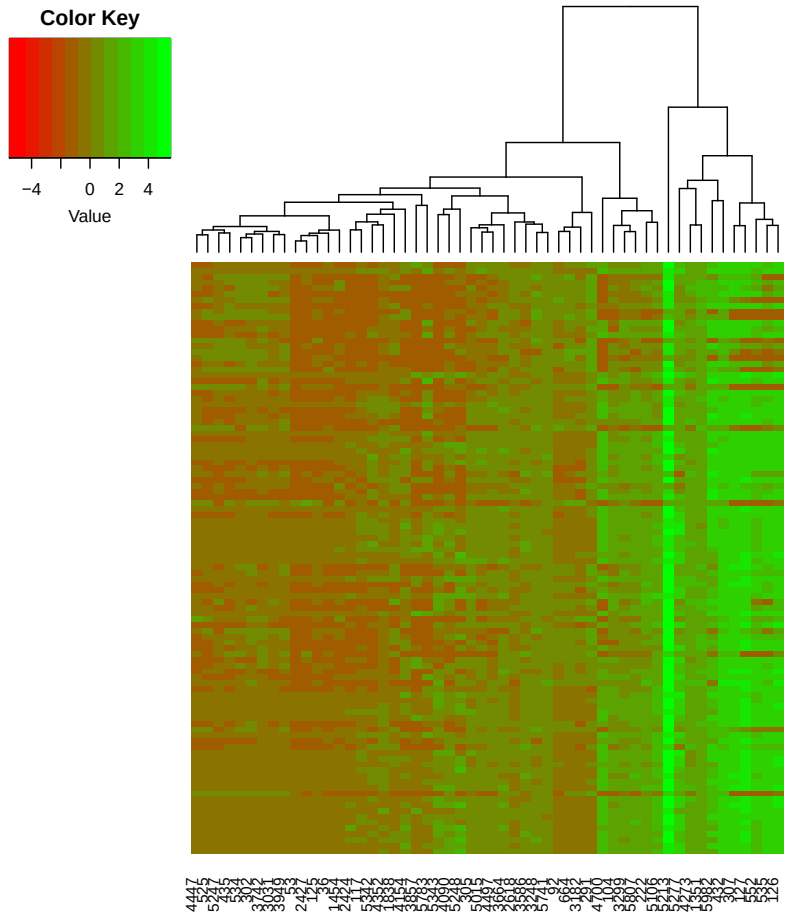
Table: SPLSDA Confusion Matrix using $\eta = 0.8$ and $K = 3$.

	Truth	
Pred	0	1
0	15	2
1	0	15

This is a much better classification compared with LDA on the same dataset. In addition, we only need to use 54 genes compared with LDA 6033 genes for some early diagnostics.

Heatmap on the Selected Genes

Figure: 54 selected genes were plotted and clustered nicely. Why ?



Nearest Shrunken Centroid Classification

- It is based on Nearest Centroid method which takes the gene expression profile of a new sample, and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that new sample.
- NSCC = NCC + shrinkage $\rightarrow$ Less gene needed.
- Detailed theory and derivation omitted here, because Xin Ma did a great job last time.

Choice of Shreshold

As well know, this shrinkage involves soft threshold as well. We present the method we select the best threshold for this data as follows,

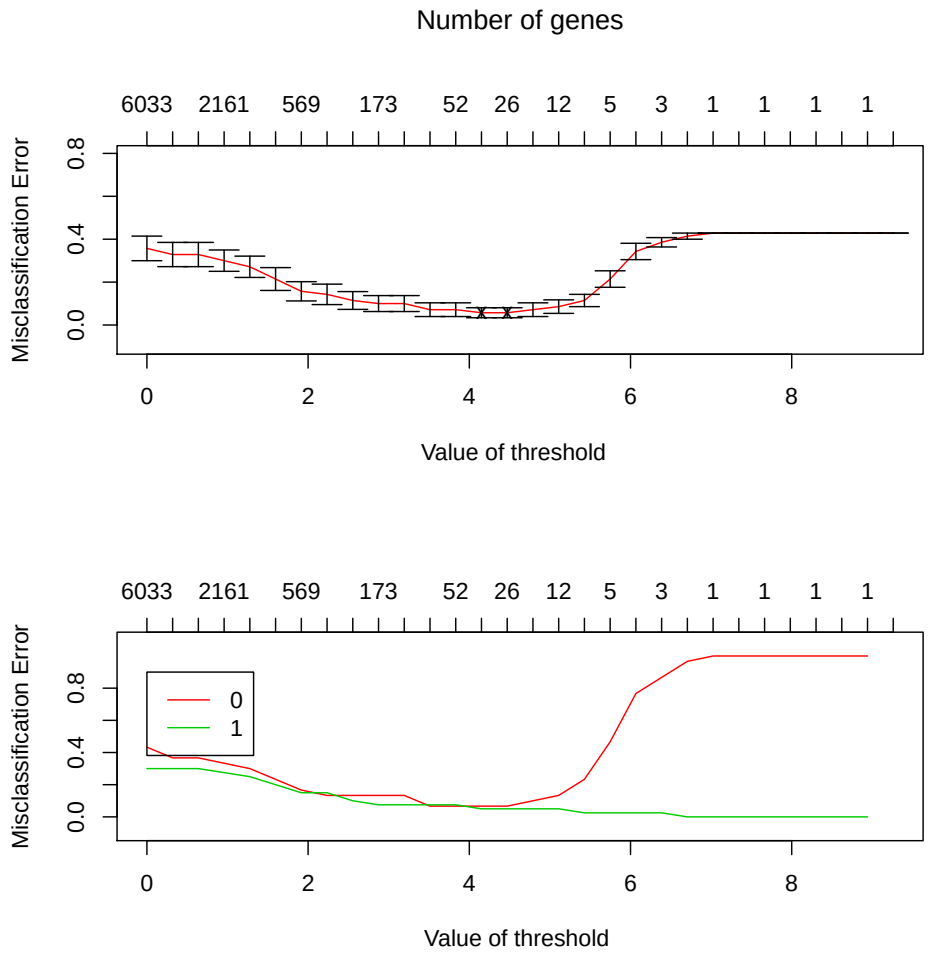
1: Propose 60 thresholds Δ from 0 to 10 and perform 10-fold cross validation.

2: Select those thresholds which produce least amount of errors and eliminate large amount of genes.

3: Use test data to finalize our choice of threshold.

Choice of Shreshold Cont.

Figure: Choice of Thresholds. How to choose from candidates ?



NSCC on Prostate Cancer Result

Table: NSCC Confusion Matrix using $\Delta = 4.43$.

	Truth	
Pred	0	1
0	13	2
1	7	10

Accuracy is only about 0.72, but different from LDA, we only need 10 genes to predict a cancer status.

Is the Above Results by Chance ?

We perform a simple simulation test using selected thresholds to build classifier on training data and then predict using testing data. To be fair, each time we partition data into 70/32 size and perform LDA, SPLSDA and NSCC.

Table: 1000 Simulated results summarized in the table.

Method	Average Misclassification	Error Rate	95% Interval
LDA	7.5 / 32	0.2344	(0.064, 0.45)
SPLS-DA	2.42 / 32	0.07563	(0, 0.17)
NSCC	6.03 /32	0.1883	(0.16, 0.25)


Show is over
YOU MAY NOW RELAX!