Class Discovery and Class Prediction by Gene Expression Monitoring

Depeng Xu

Xiahan Tang

James Willbanks

Md Kamrul Hasan Khan

OUTLINE

- Introduction
- Data Description
- Data Preprocessing
- Methods
  - Unsupervised
  - Supervised

INTRODUCTION

We analyzed gene expression data using unsupervised and supervised methods to find possible ways to classify if a leukemia patient has acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML).

DATA DESCRIPTION

- Data retrieved from a paper by Golub et al. in 1999
- 72 samples divided into 38 training and 34 testing points
- Each sample has 7129 gene levels
- Binary response
- ALL = 0 & AML = 1

DATA PREPROCESSING

- Dudoit, S., Fridlayand, J. and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumor Using Gene Expression Data", *JASA*.
- Data processed by:

  Thresholding: thresholding with floor of 100 and ceiling of 16000

  Filtering: excluding genes with max/min ≤ 5 or (max – min) ≤ 500

  Transformation: transforming by base 10 logarithm

  Standardization: standardize so that the observations have mean 0 and variance 1 across the variables (genes)

- 3051 Genes
- 263 Genes

METHODS

- **Unsupervised:** k-means, KNN, Decision Tree, Hierarchical Clustering, PCA
- **Supervised:** PCR, Regression

UNSUPERVISED METHODS

K-means

K = 2	Actual classifications		
Predicted Classification		ALL	AML
	ALL	25	0
	AML	2	11

K = 3	Actual classifications		
Predicted Classification		ALL	AML
	ALL	26	0
	AML	1	11

K = 4	Actual classifications		
Predicted Classification		ALL	AML
	ALL	27	0
	AML	0	11

K-Nearest Neighbor

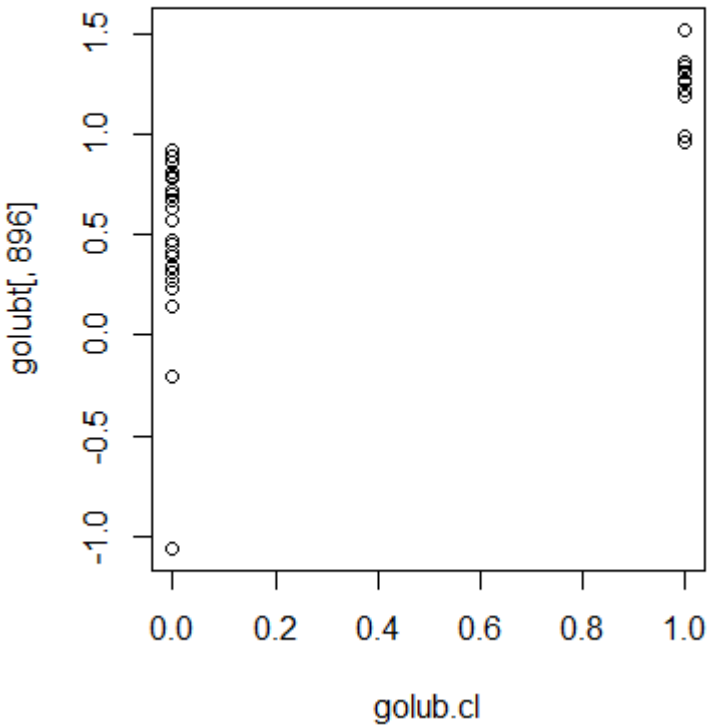
K = 3		Actual Level	
		ALL	AML
Cluster Prediction	ALL	26	0
	AML	1	11

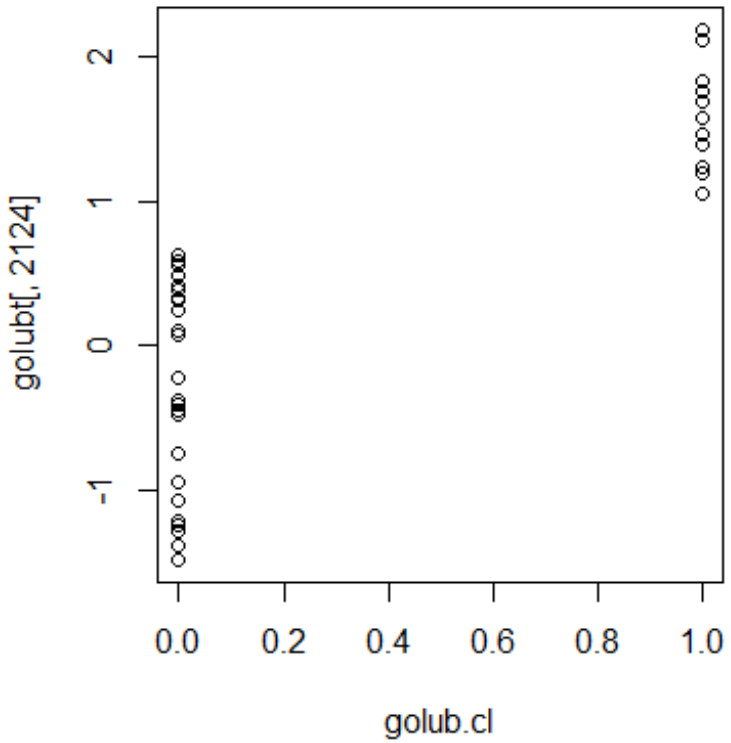
K = 4		Actual Level	
		ALL	AML
Cluster Prediction	ALL	26	0
	AML	1	11

K = 5		Actual Level	
		ALL	AML
Cluster Prediction	ALL	26	0
	AML	1	11

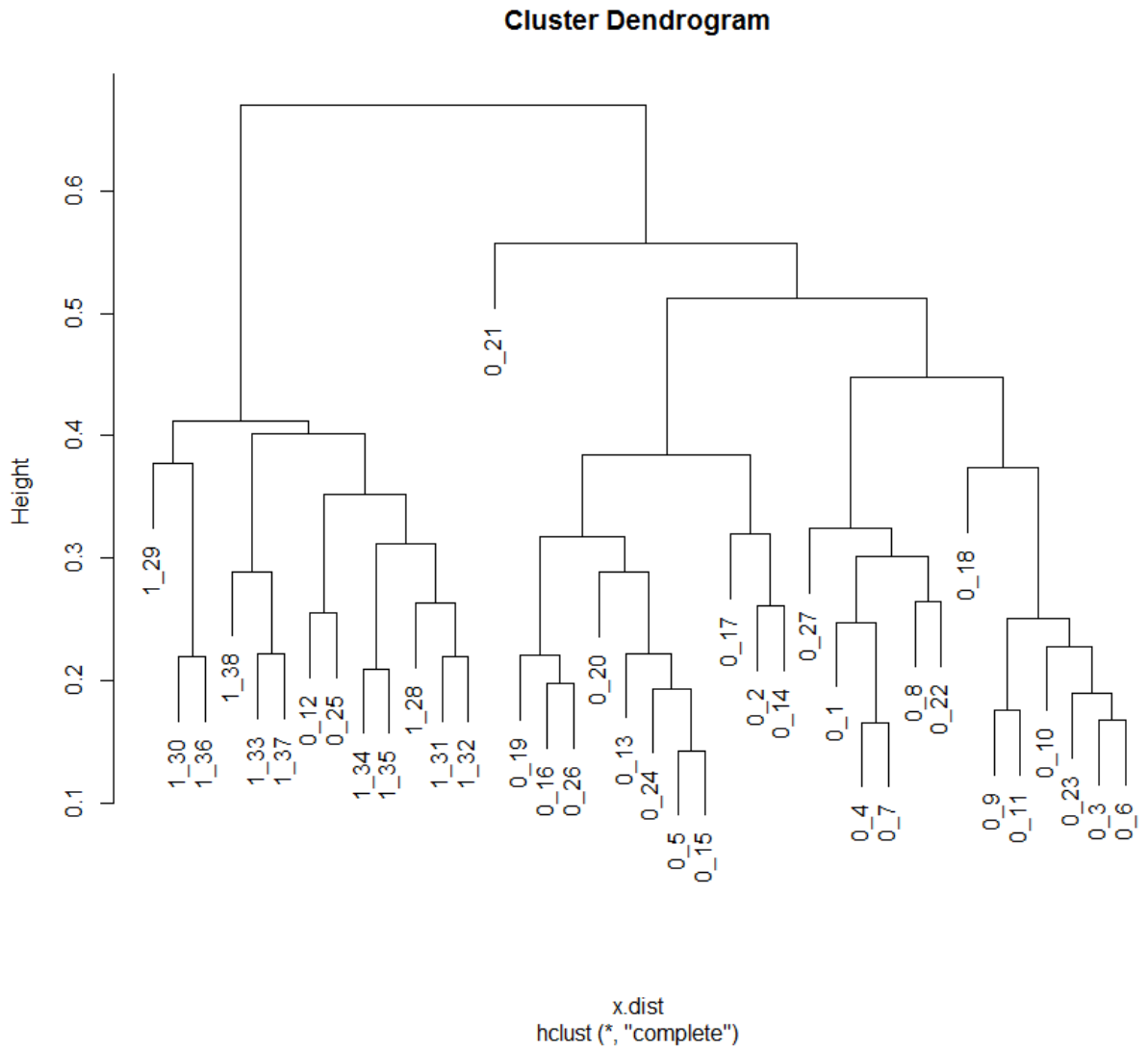
K = 6		Actual Level	
		ALL	AML
Cluster Prediction	ALL	27	0
	AML	0	11

Decision Trees



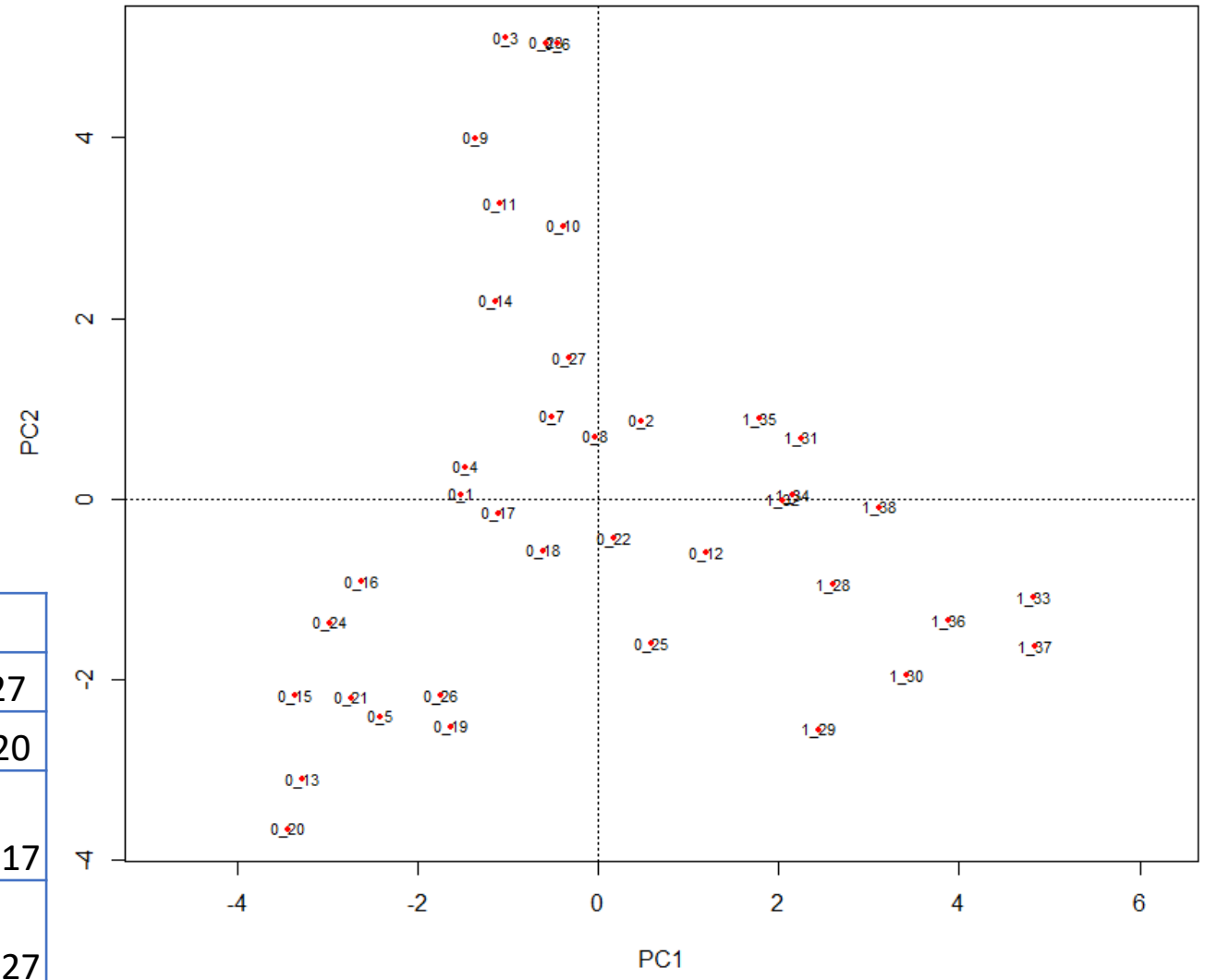


HIERARCHICAL CLUSTERING



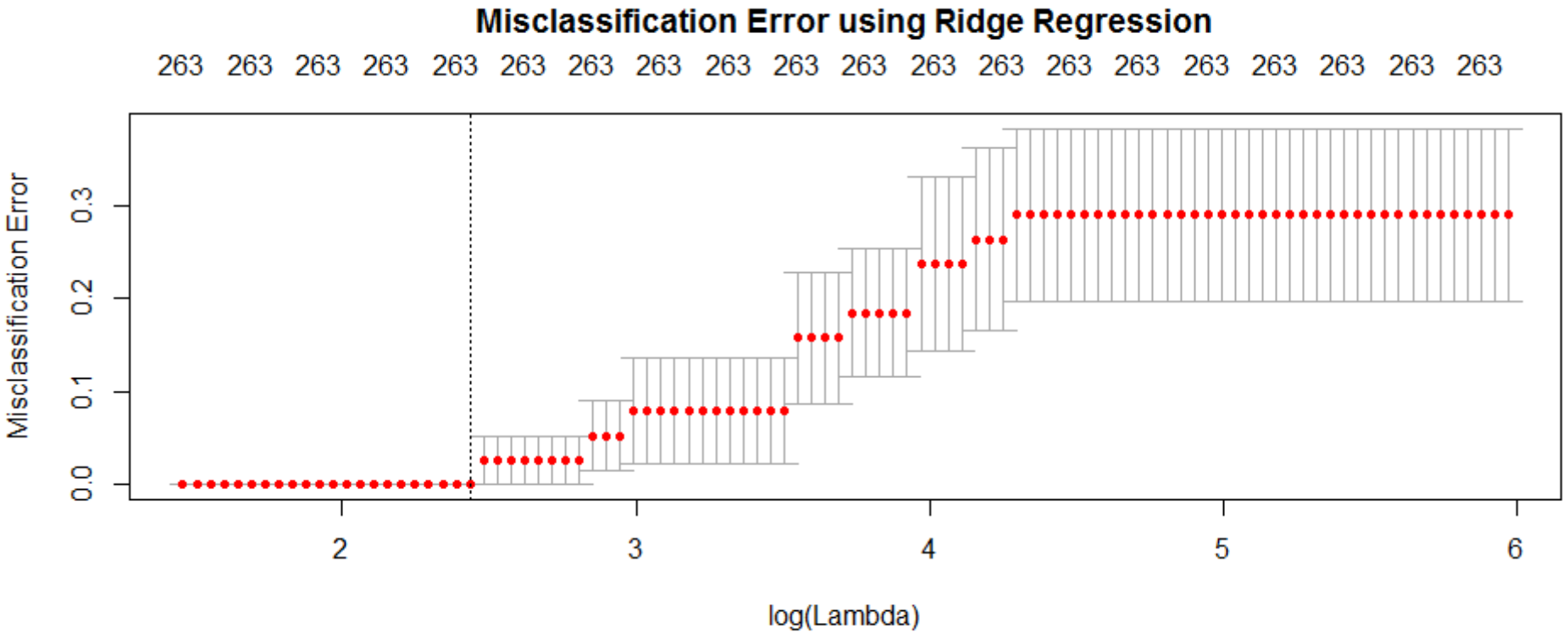
PCA

Eigenvalues, and their contribution to the variance							
	PC1	PC2	PC3	PC4	PC5	…	PC27
Eigenvalue	171.44	103.52	88.43	62.43	46.60	…	12.20
Proportion Explained	0.1645	0.0993	0.0849	0.0599	0.0447	…	0.0117
Cumulative Proportion	0.1645	0.2639	0.3487	0.4086	0.4533	…	0.9027



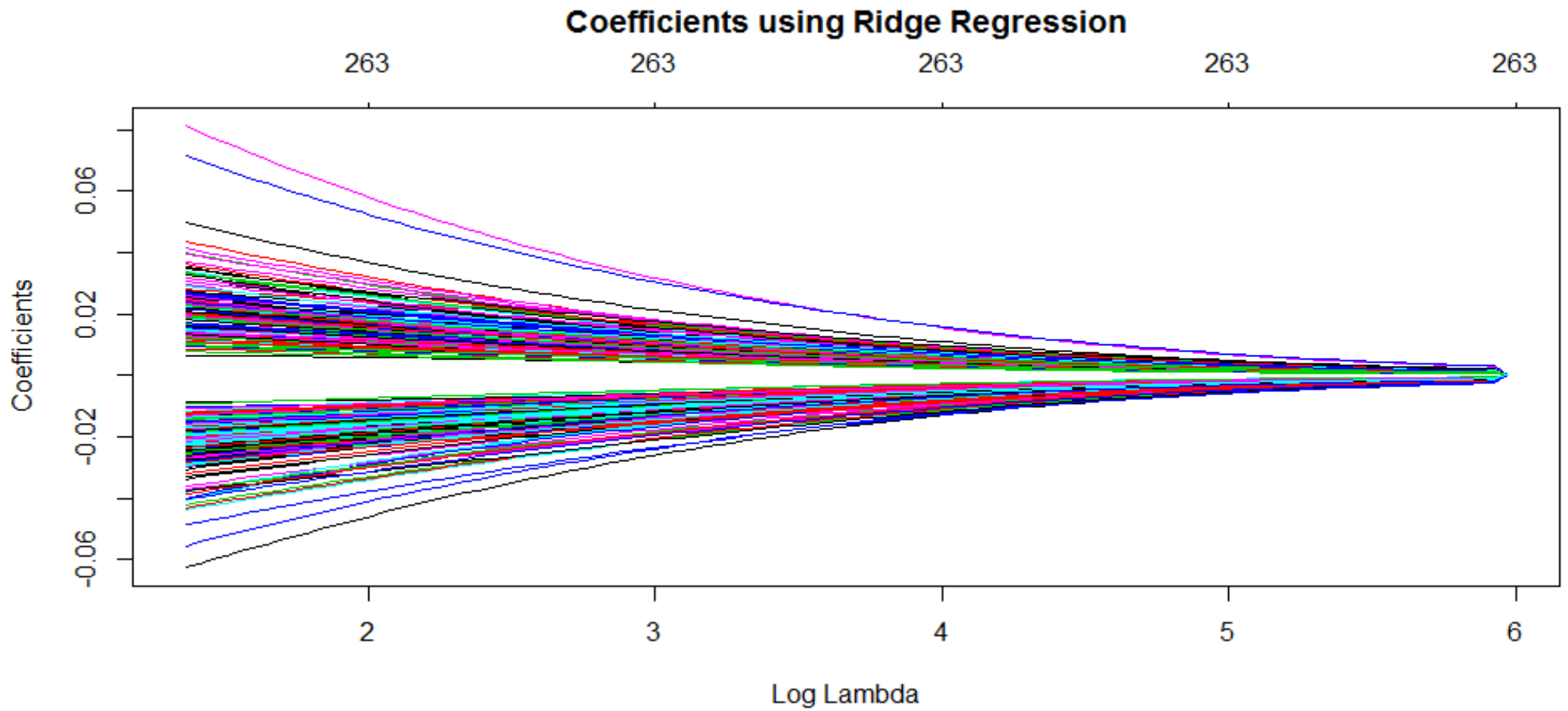
SUPERVISED METHODS

RIDGE REGRESSION



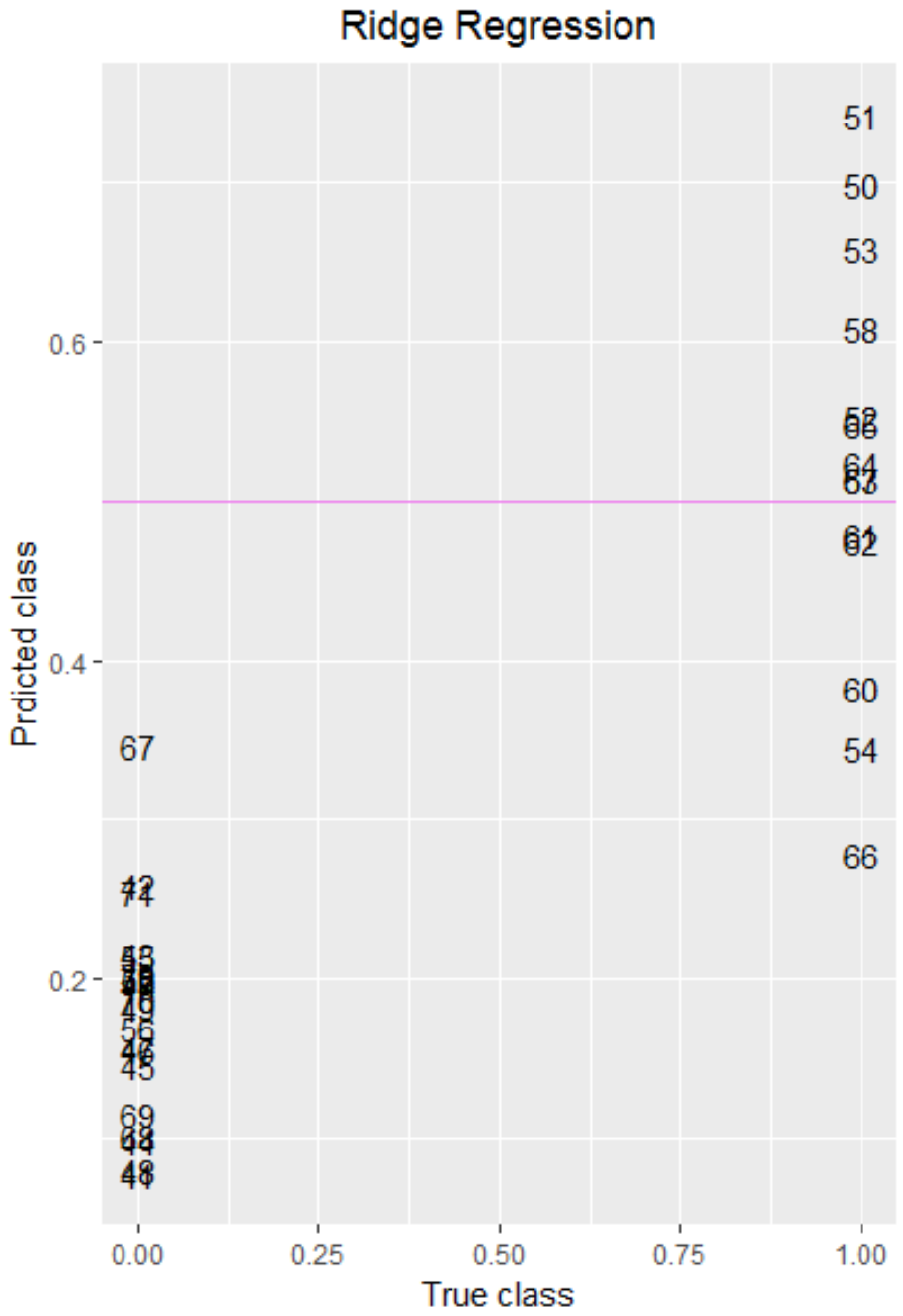
Best Lambda = 17.34371

RIDGE REGRESSION (CTD.)

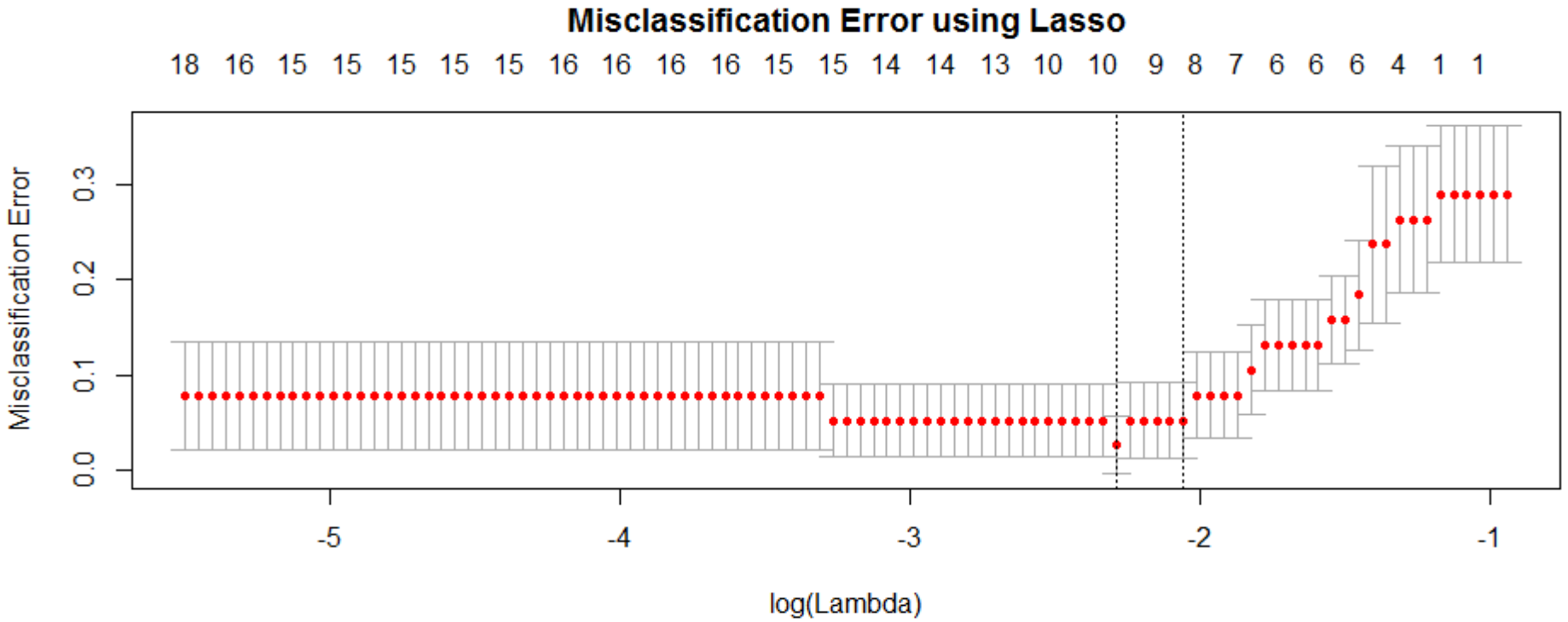


RIDGE REGRESSION(CTD.)

		True Class	
		0	1
Predicted Class	0	20	5
	1	0	9



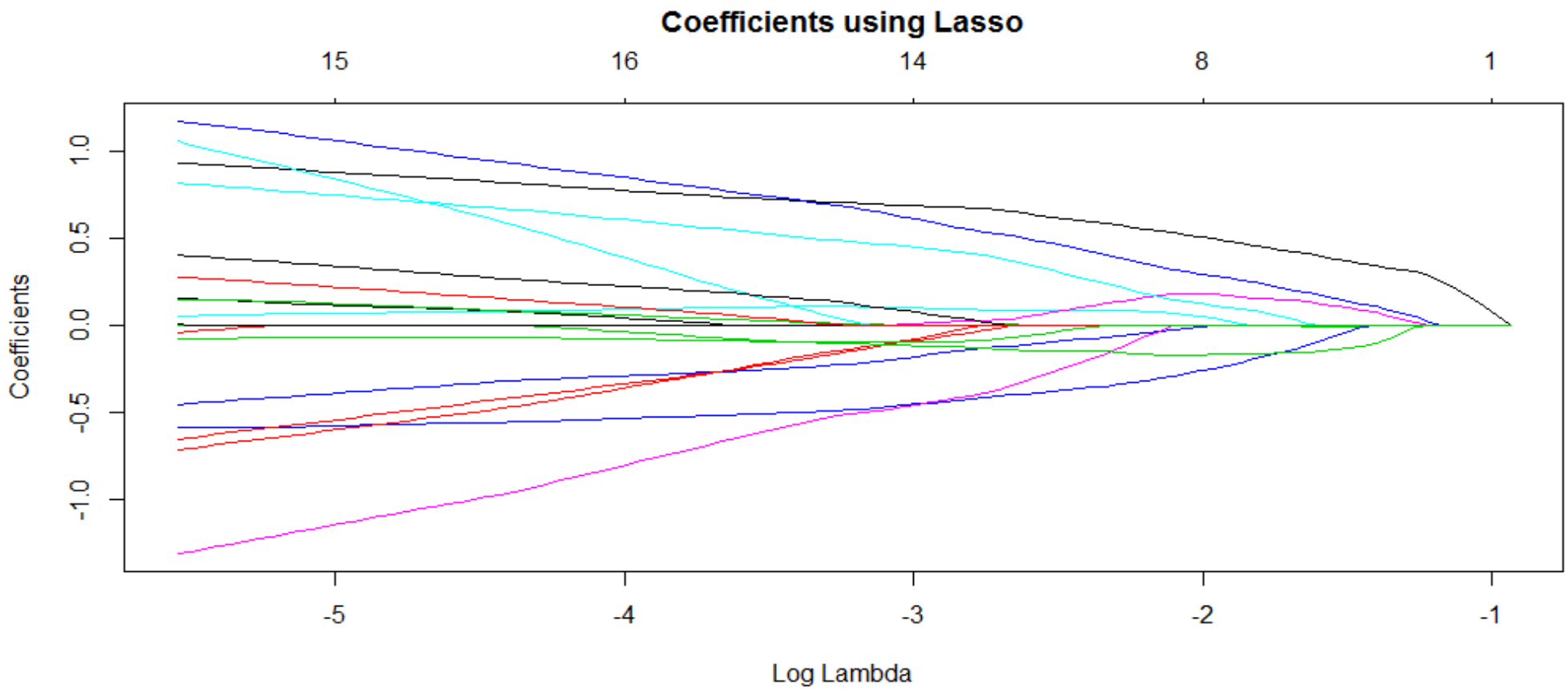
LASSO REGRESSION



Lambda.1se = 0.1281825

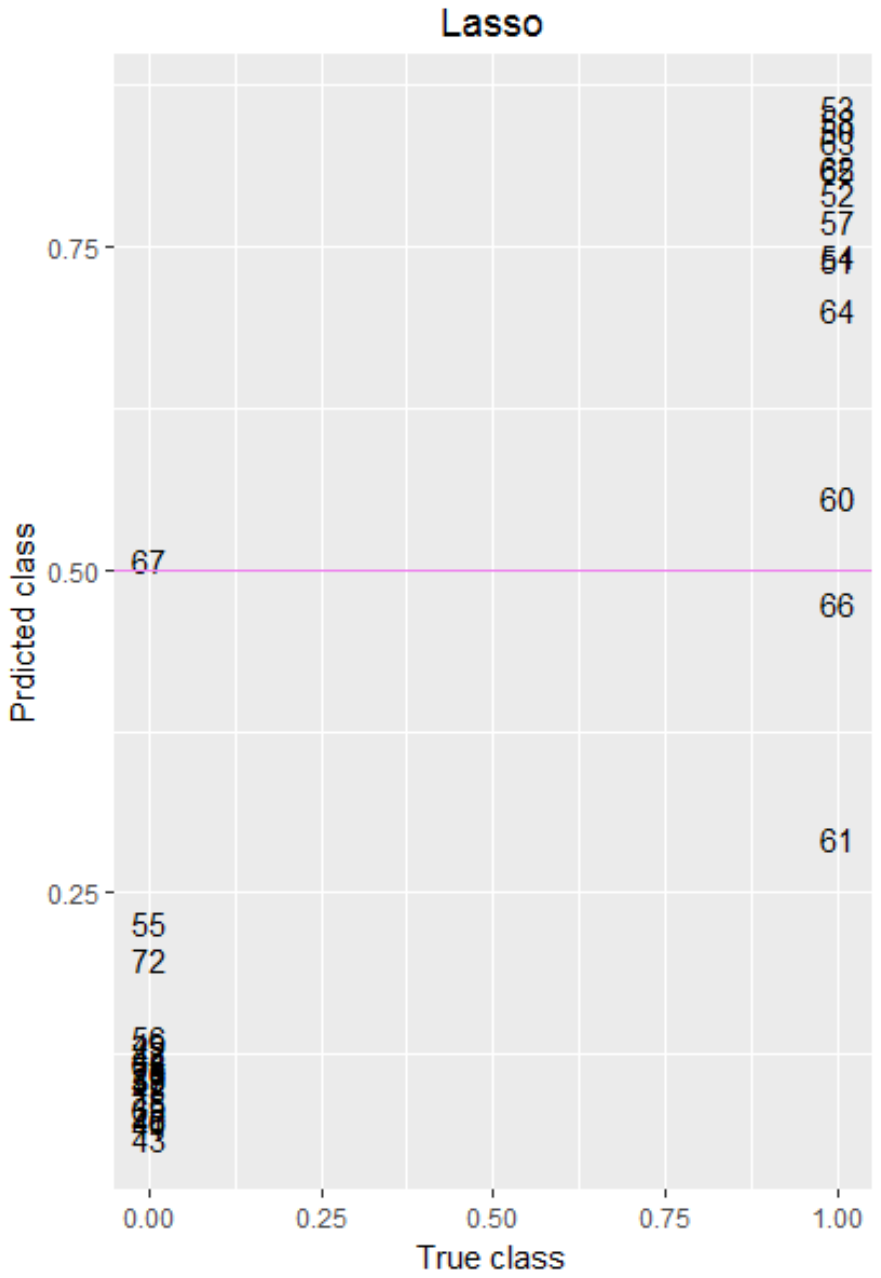
Lambda.min = 0.1015824

LASSO REGRESSION (CTD.)

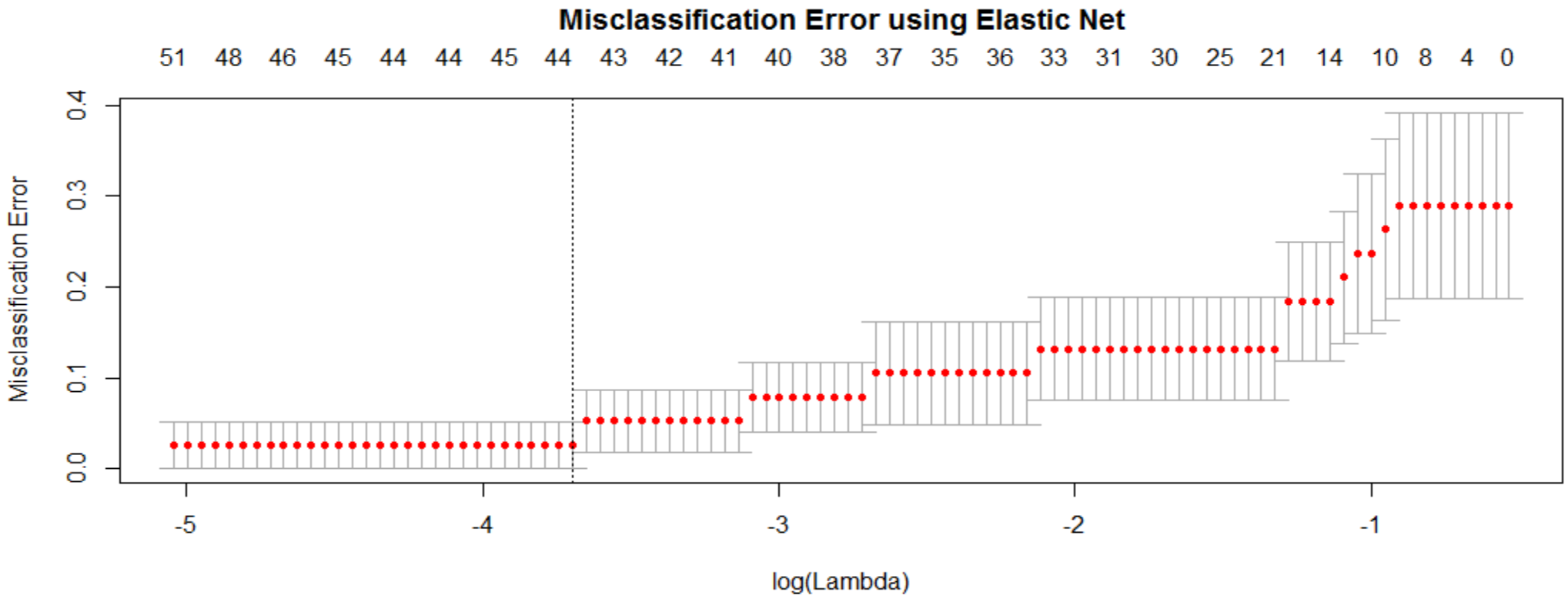


LASSO REGRESSION (CTD.)

		True Class	
		0	1
Predicted Class	0	19	2
	1	1	12



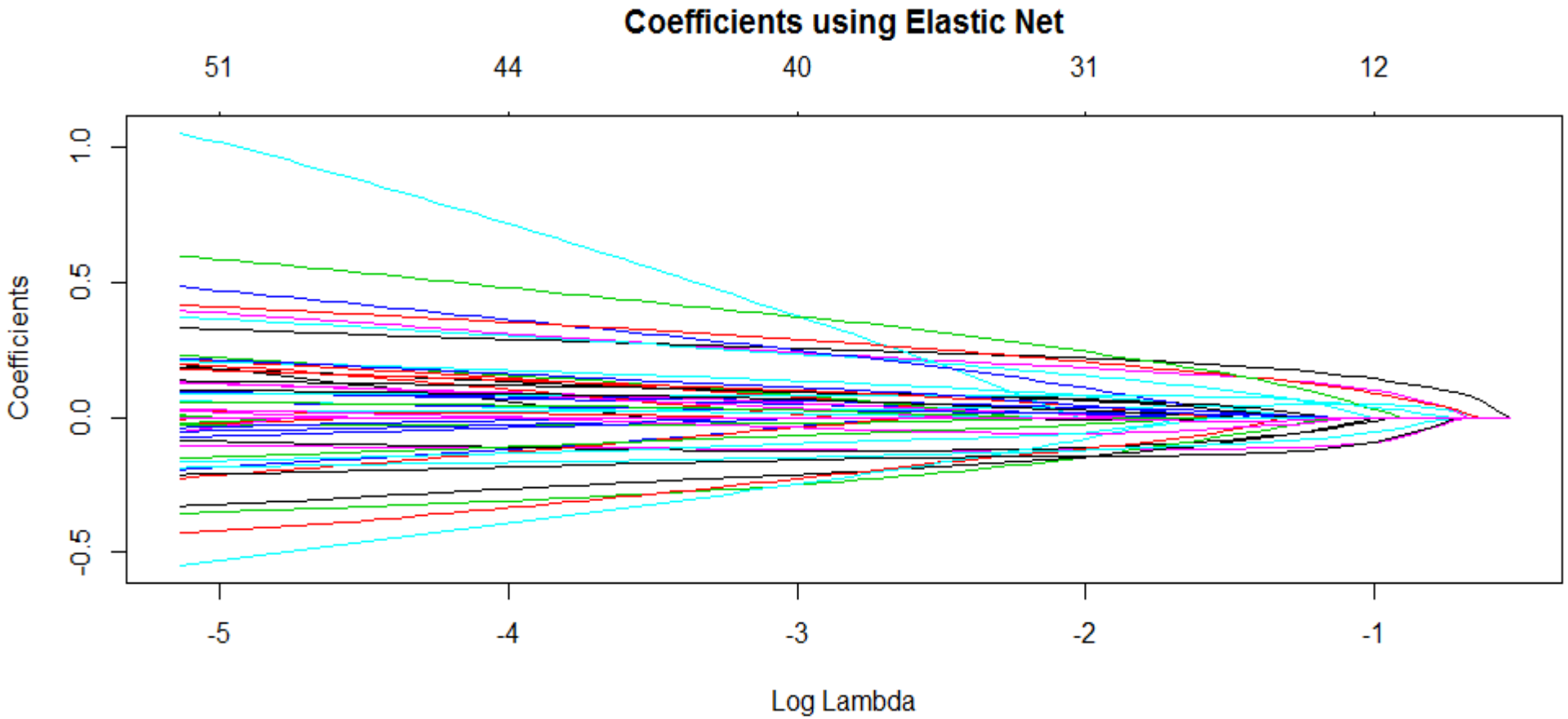
ELASTIC NET



Best α = 0.66666667

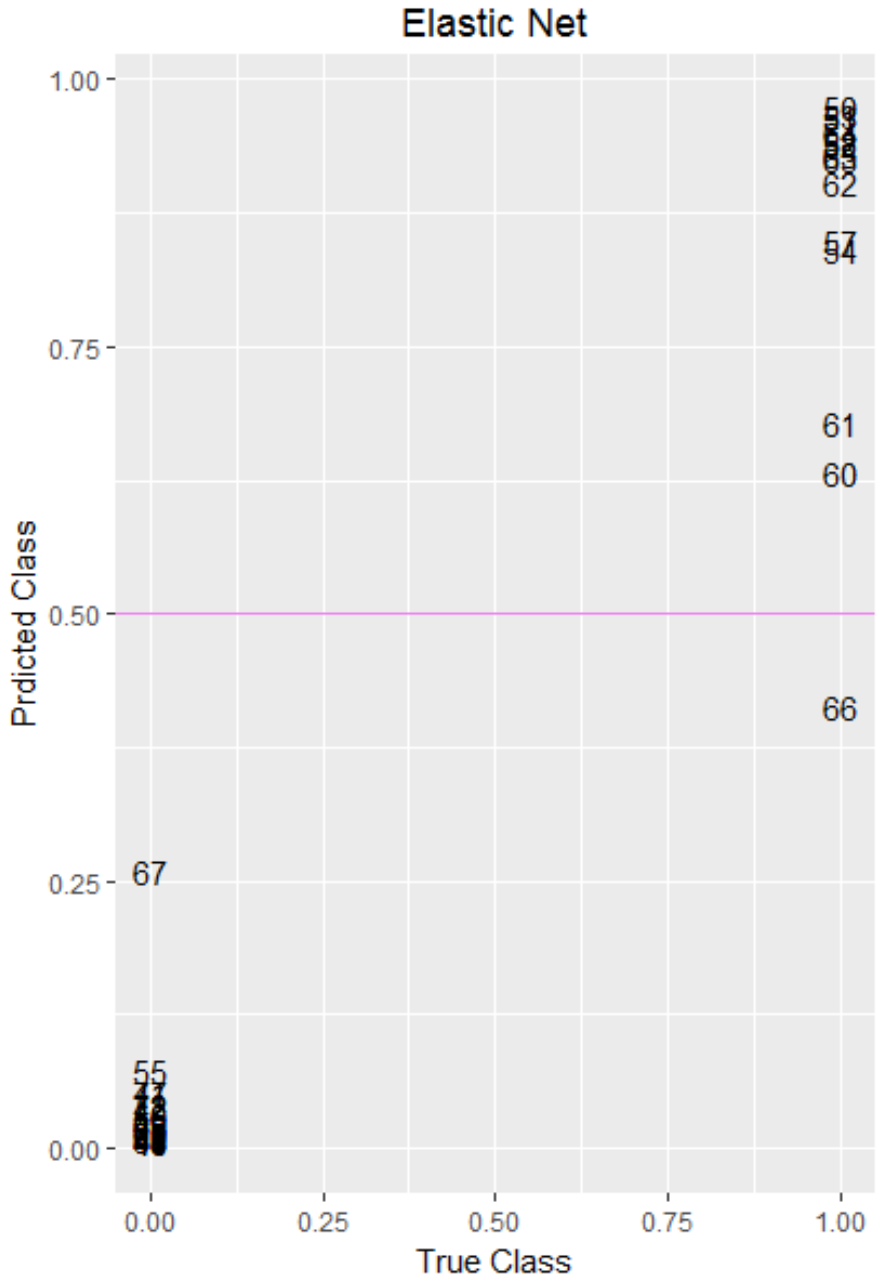
Lambda.1se = 0.02483311

ELASTIC NET (CTD.)



ELASTIC NET (CTD.)

		True Class	
		ALL	AML
Predicted Class	ALL	20	1
	AML	0	13

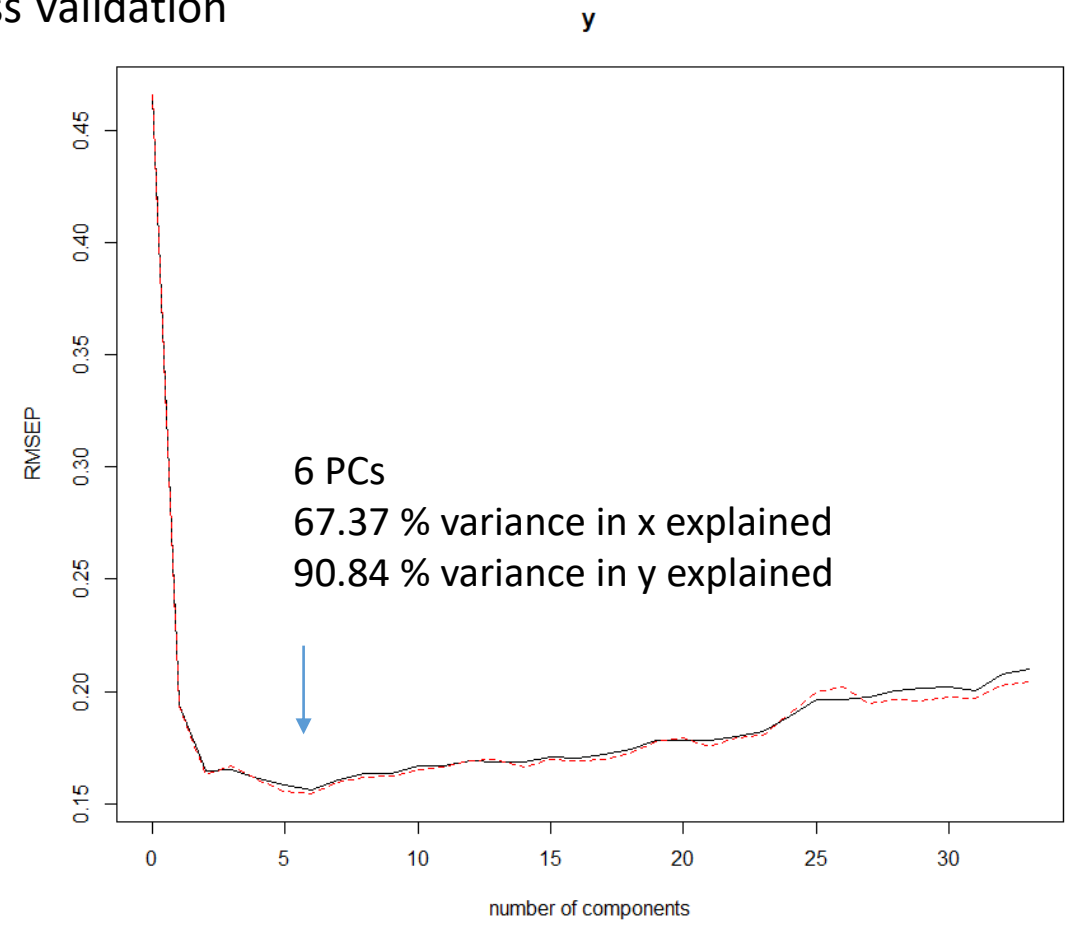


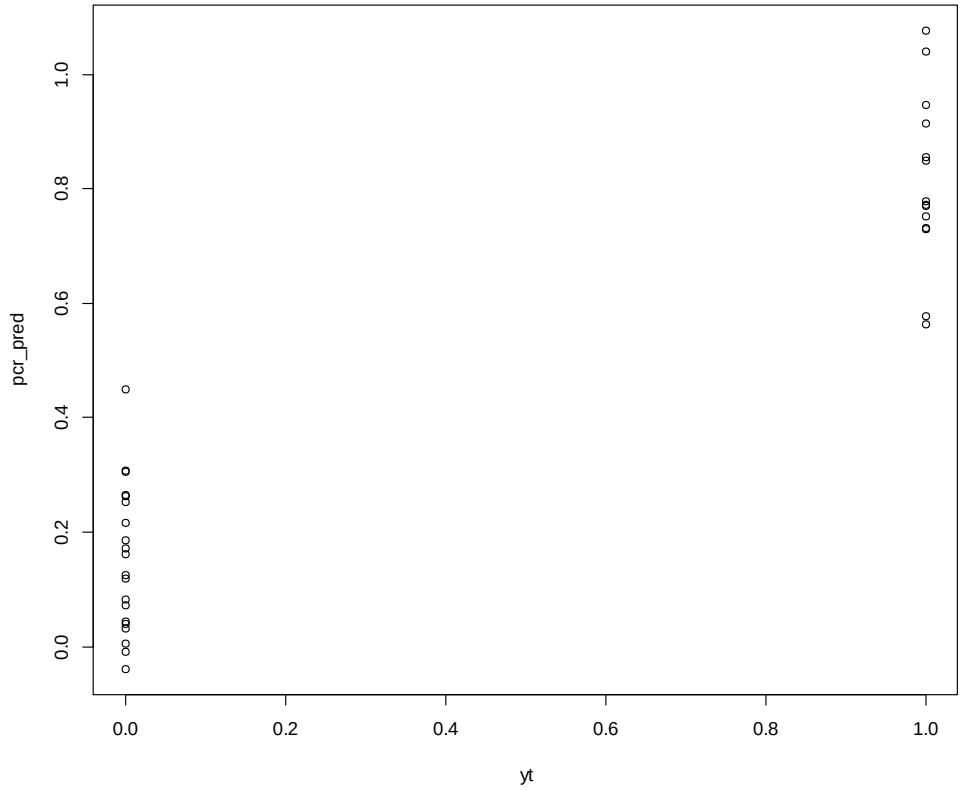
Variable Selection using Ridge, Lasso and Elastic Net

	Number of Non-zero Beta
Ridge	263
Lasso	8
Elastic Net	44

PCR

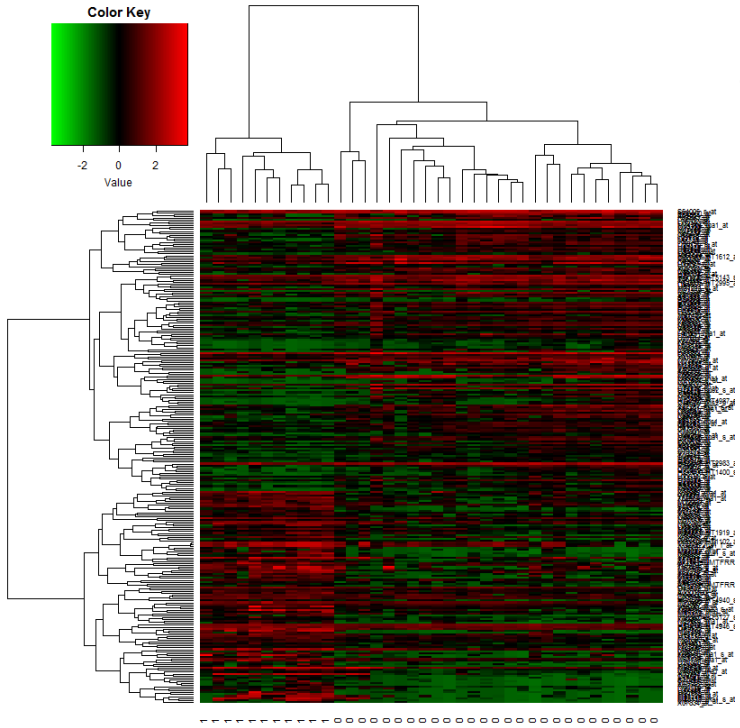
Cross Validation



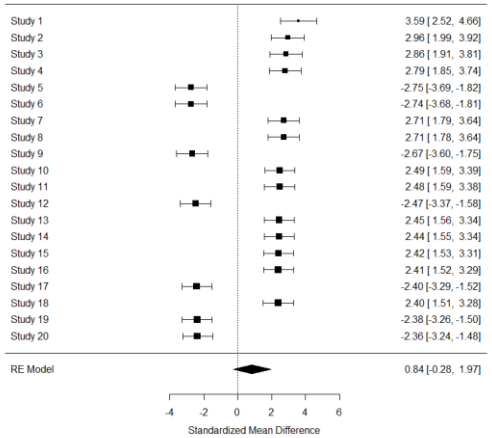


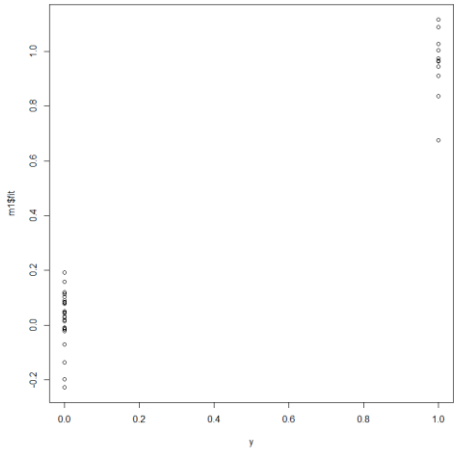
		True Class	
		0	1
Predicted Class	0	20	0
	1	0	14

Others- Characteristic genes

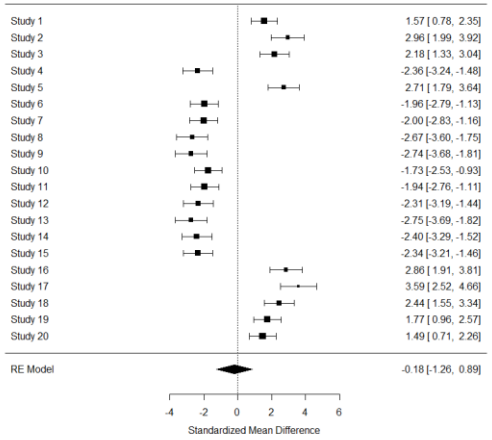


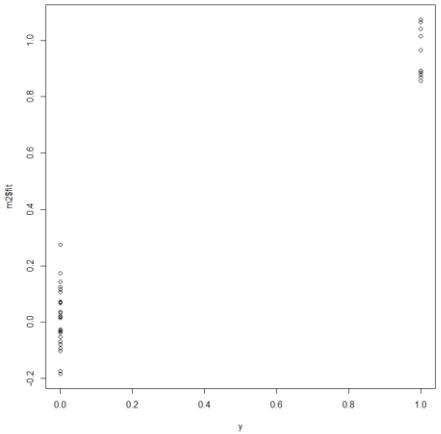
Top 20 genes





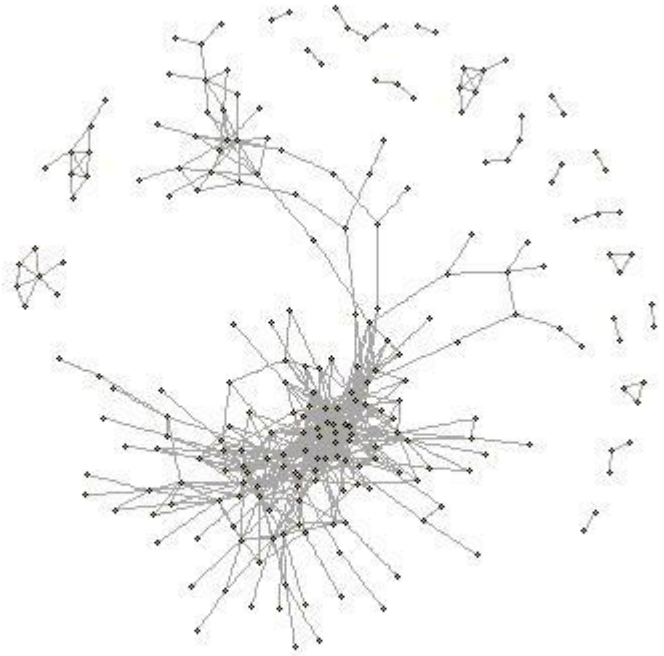
Characteristic genes
of 20 gene groups





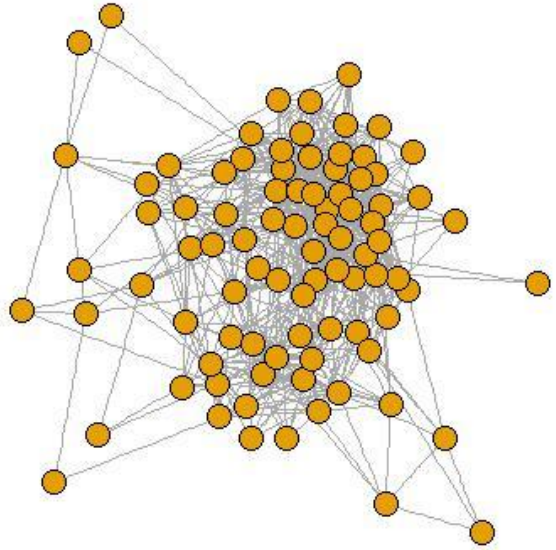
Other- Network Analysis

acute lymphoblastic leukemia (ALL)



Cutoff=0.68, #nodes=234, #links=640

acute myeloid leukemia (AML)



Cutoff=0.56, #nodes=88, #links=640

Questions?