Molecular classification of cancer: a PCA approach

Xinyang Liu, Jinghan Cui

4/25/2022

1 Introduction

The project is inspired by the study published in 1999 by Golub *et al* in which researchers developed a systematic approach to cancer classification based on global gene expression analysis using DNA microarrays. Traditionally, diagnosis of cancer has relied on histopathological appearance, but a serious limitation is that tumors with similar histopathological appearance may have different clinical courses and responses to therapy.

Taking acute leukemias for example, there are two sub-types: acute lymphoblastic leukemia(ALL) and acute myeloid leukemia(AML). It is important to distinguish ALL from AML for target treatment. The distinction between them can be well done in clinical practice, but misclassification may occur sometimes.

We developed a classification model based on the gene expression data. Out of thousands of genes, we tried to identify a small portion of gene with significantly different gene expression levels. If we have a new, unknown sample of acute leukemia, then researchers can perform gene sequencing on targeted gene we have identified and use the classification model as an assistance to the diagnosis of cancer sub-type.

2 Data

We used the same data as Golub *et al* used in their study. The dataset consists of quantitative expression levels of 7192 genes from 72 acute leukemia patients. The patients are labeled with Acute myeloid leukemia(AML) and Acute lymphoblastic leukemia(ALL) from previous clinical diagnoses.

3 Methods

Since this is a problem of "$p >> n$", dimension reduction should be done before we implement a logistic regression model. Hence, we will perform Principal Component Analysis (PCA) on the gene expression data for dimension reduction. Then we use the principal components to classify the types of leukemia. Comparing the accuracies of classification with different numbers of components, we pick the model with the best balance of sensitivity and specificity. Next, we calculate the bootstrap confidence interval for each component along with the z-score confidence interval and find the significant one. We then analyze the components using marginal correlation to identify genes that are differentially expressed between AML and ALL.

Contribution All authors contributed equally to each part of the project including data analysis, writing, and presenting.

Reference

Golub, T. R., et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." Science, vol. 286, no. 5439, 1999, pp. 531–37, http://www.jstor.org/stable/2899325. Accessed 1 May 2022.

Codes and lecture notes from UMass Amherst Spring 2022 STAT 697MV course by Professor Shai Gorsky.

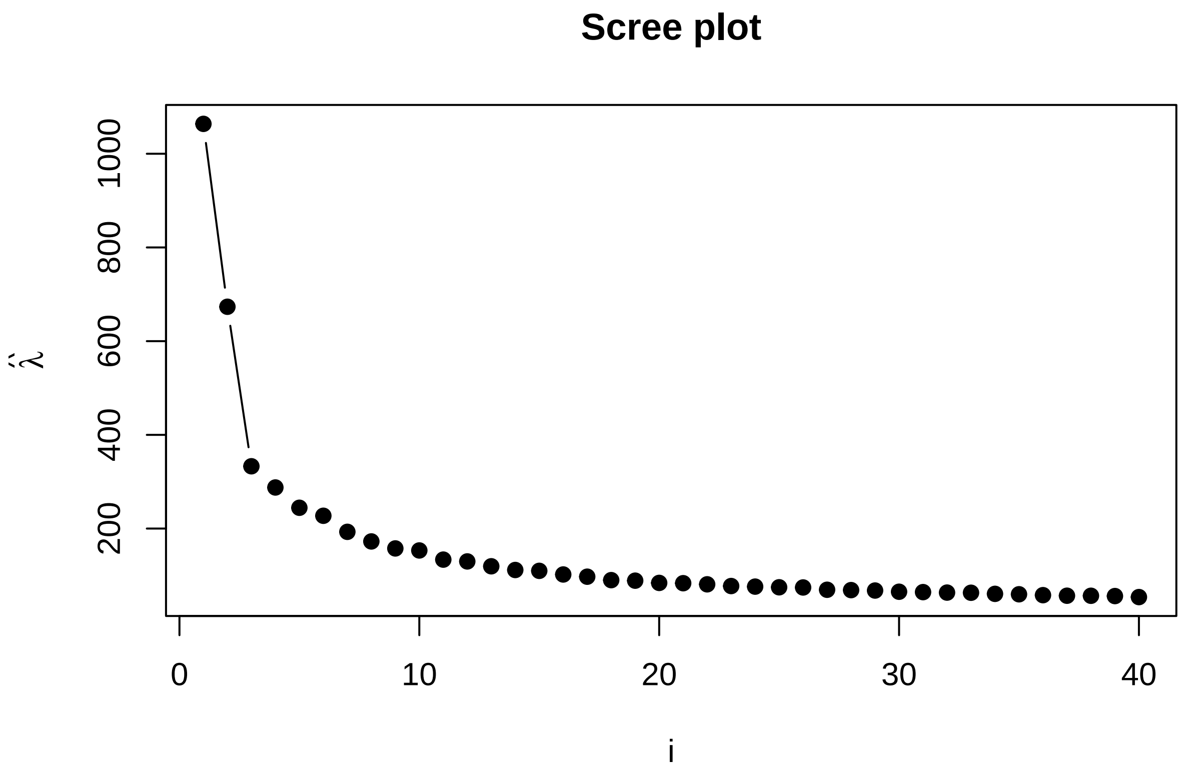
4 Data analysis

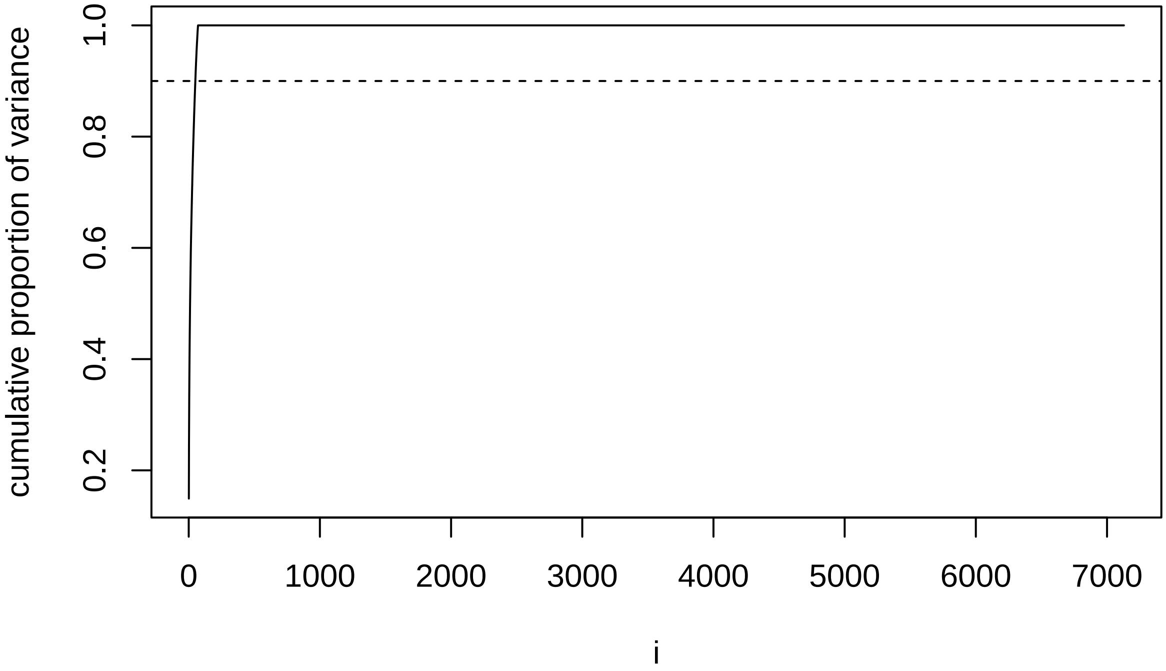
First we did a data cleaning and since the gene expressions ranges differently, we did a standardization.

Next, we calculate the covariance matrix, eigenvalues and eigenvector.

```
##   [1] 0.1492159 0.2436776 0.2903799 0.3307340 0.3650138 0.3968978 0.4239678
##   [8] 0.4481667 0.4702742 0.4917734 0.5105290 0.5287436 0.5454990 0.5611507
##  [15] 0.5765459 0.5908498 0.6044965 0.6171056 0.6295613 0.6413423 0.6530387
##  [22] 0.6643933 0.6752439 0.6859152 0.6963878 0.7067999 0.7165289 0.7261552
##  [29] 0.7356536 0.7447891 0.7538111 0.7626899 0.7715230 0.7800292 0.7884114
##  [36] 0.7965295 0.8044770 0.8124095 0.8202519 0.8277941 0.8352131 0.8425064
##  [43] 0.8497479 0.8567759 0.8637284 0.8706090 0.8774060 0.8841694 0.8907413
##  [50] 0.8972340 0.9034596 0.9095273 0.9155376 0.9213419 0.9271066 0.9328108
##  [57] 0.9383162 0.9438018 0.9490414 0.9542160 0.9591344 0.9639833 0.9685209
##  [64] 0.9728142 0.9770745 0.9812341 0.9853334 0.9893166 0.9930381 0.9966236
##  [71] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [78] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [85] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [92] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [99] 1.0000000 1.0000000
```

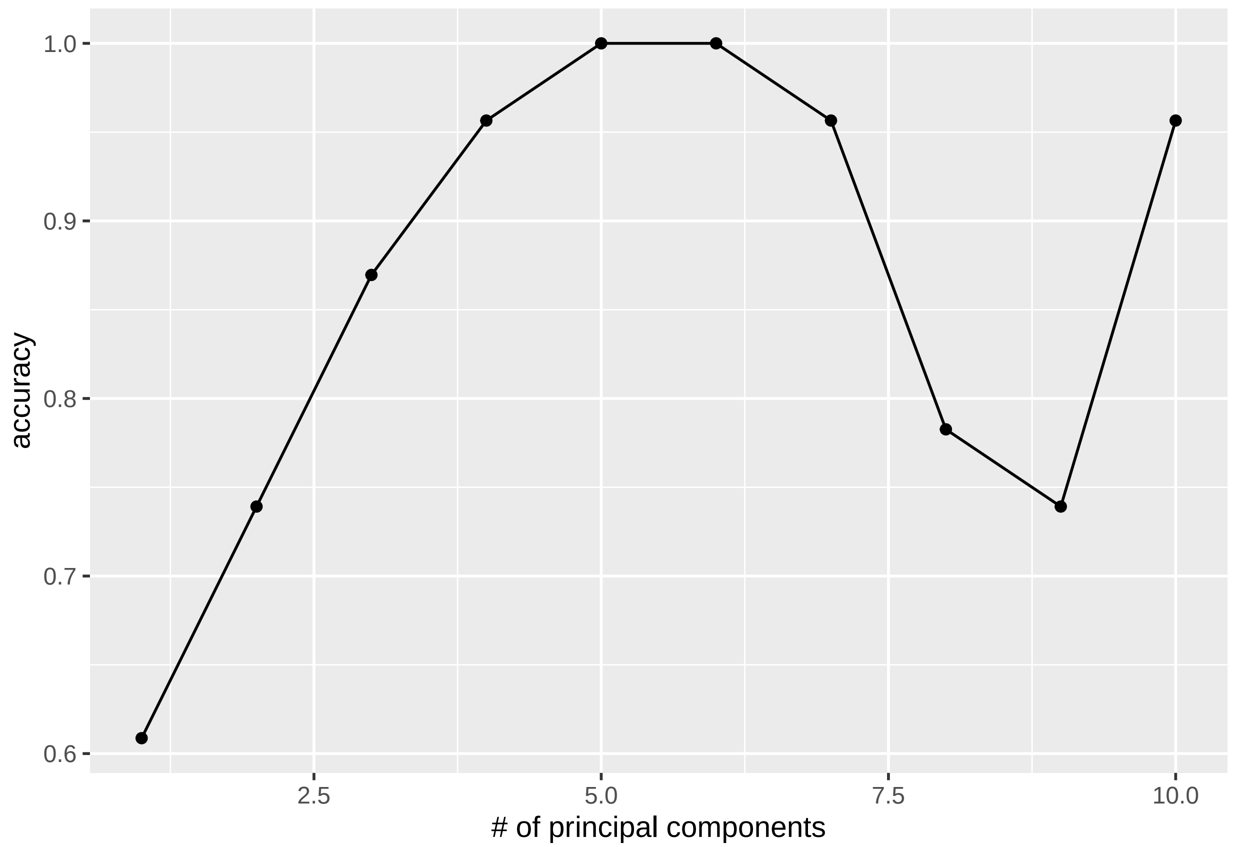
From the cumulative variance and scree plot, we found 70 components can explain 99% of the covariance.





Due to the limitation of computing power, we decided to run logistic regression for 1 to 10 component. 10 component can explain approximate 50% of the variance.

The data is splited into train and test dataset at a ratio of 1:2. The accuracy of classification in the test dataset is calculated.



From the accuracy plot we can see the accuracy reaches 100% when we include 5 principal components to the model and then decreases as we include more components but increases again to 95% when we include 10 components. We think this unstable results may be caused by the small sample size because when we include more than 5 components to the model, the model can not converge.

We chosed the model with three principal components as the 3-component model can converge and has an accuracy of 86%. This model should have a good balance between bias and variance. We want to run a bootstrap for this model and check whether the regression 95% confidence interval and the bootstrap 95% confidence interval agrees.

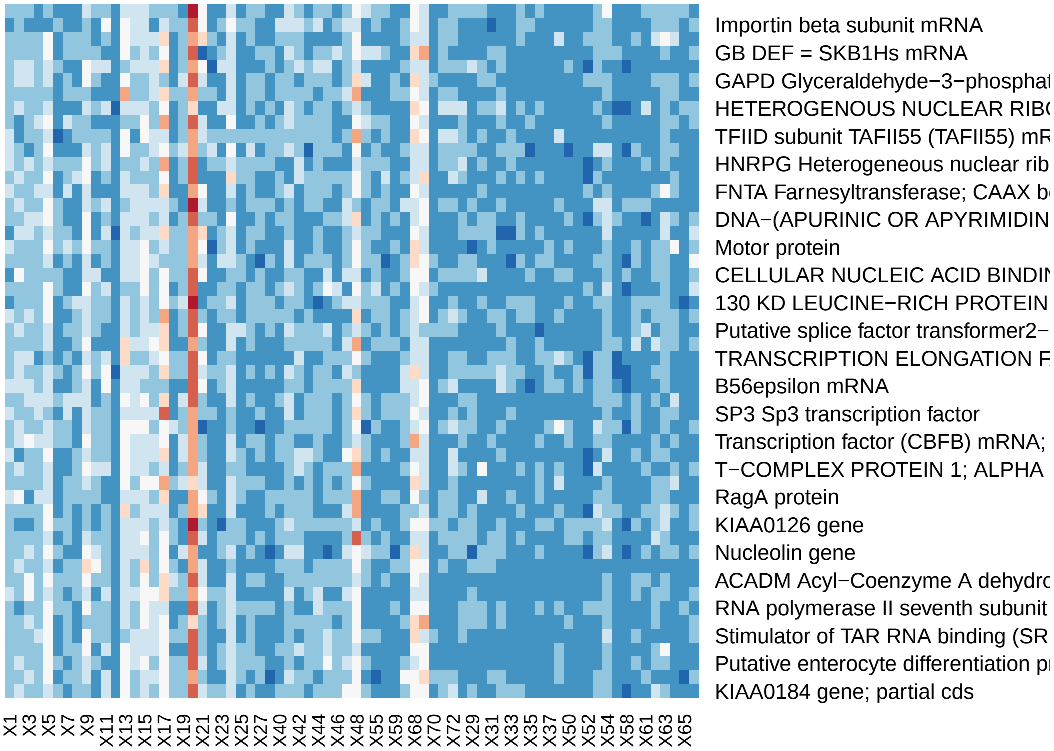
	2.5 %	97.5 %	2.5 %	97.5 %
	regression	regression	bootstrap	bootstrap
(Intercept)	0.4575	3.3251	0.5693	5.2647
V1	-0.0695	6e-04	-0.0856	4e-04
V2	-0.2169	-0.0388	-0.442	-0.0741
V3	0.0588	0.2362	-0.4418	-0.0737

Both the regression and boostrap confidence intervals agree that the first component is not significant and the second component is significantly negatively correlated with the outcome. While for the third component, the regression confidence interval shows it is positively correlated but the bootsrap confidence interval show it is negatively correlated. So we will further explore the second component.

We want to look into it's marginal correlations to determine which gene expression contributes most to this component by calculating the marginal correlations.

We picked the 50 genes that contributes the most to the second component and draw a heatmap to see how thoes genes' expression levels are related to the outcome.]

The left side of the heatmap are patients who are diagnosed with Acute myeloid leukemia(AML), and the right side are patients who are diagnosed with Acute lymphoblastic leukemia(ALL). We can see that most of the genes are much more down-regulated for ALL patients than for AML patients.


Importin beta subunit mRNA
GB DEF = SKB1Hs mRNA
GAPD Glyceraldehyde−3−phosphat
HETEROGENOUS NUCLEAR RIB
TFIID subunit TAFII55 (TAFII55) mR
HNRPG Heterogeneous nuclear rib
FNTA Farnesyltransferase; CAAX b
DNA−(APURINIC OR APYRIMIDIN
Motor protein
CELLULAR NUCLEIC ACID BINDIN
130 KD LEUCINE−RICH PROTEIN
Putative splice factor transformer2−
TRANSCRIPTION ELONGATION F
B56epsilon mRNA
SP3 Sp3 transcription factor
Transcription factor (CBFB) mRNA;
T−COMPLEX PROTEIN 1; ALPHA
RagA protein
KIAA0126 gene
Nucleolin gene
ACADM Acyl−Coenzyme A dehydro
RNA polymerase II seventh subunit
Stimulator of TAR RNA binding (SR
Putative enterocyte differentiation p
KIAA0184 gene; partial cds
X1
X3
X5
X7
X9
X11
X13
X15
X17
X19
X21
X23
X25
X27
X40
X42
X44
X46
X48
X55
X59
X68
X70
X72
X29
X31
X33
X35
X37
X50
X52
X54
X58
X61
X63
X65