Molecular classification of cancer: a PCA approach

Xinyang Liu, Jinghan Cui

5/5/2022

Content

- Motivation
- Data
- PCA approach
- Results

Motivation

- Inspired by the study published in 1999 by Golub *et al*
  - A systematic approach to cancer classification based on global gene expression analysis using DNA microarrays.
  - They proposed "neighborhood analysis" to identify genes that are more highly correlated with the tumor class distinction
- Example: Acute leukemias
  - Two sub-types: acute lymphoblastic leukemia(ALL) and acute myeloid leukemia(AML)
  - Important to distinguish ALL from AML for target treatment.

Motivation

- Why using gene expression data that may introduce more variation?
  - The distinction between them has been well established in clinical practice by experienced hematopathologist. But misclassification may occur sometimes.
  - 1) Develop a classification model based on the gene expression data
  - 2) Identify a small portion of gene with significantly different gene expression levels
  - 3) given a new sample of acute leukemia, 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

Data

- The dataset consists of quantitative expression levels of $p = 7192$ genes from $n = 72$ acute leukemia patients.
- Out of 72 patients, 47 have Acute lymphoblastic leukemia(ALL) and 25 have Acute myeloid leukemia(AML).

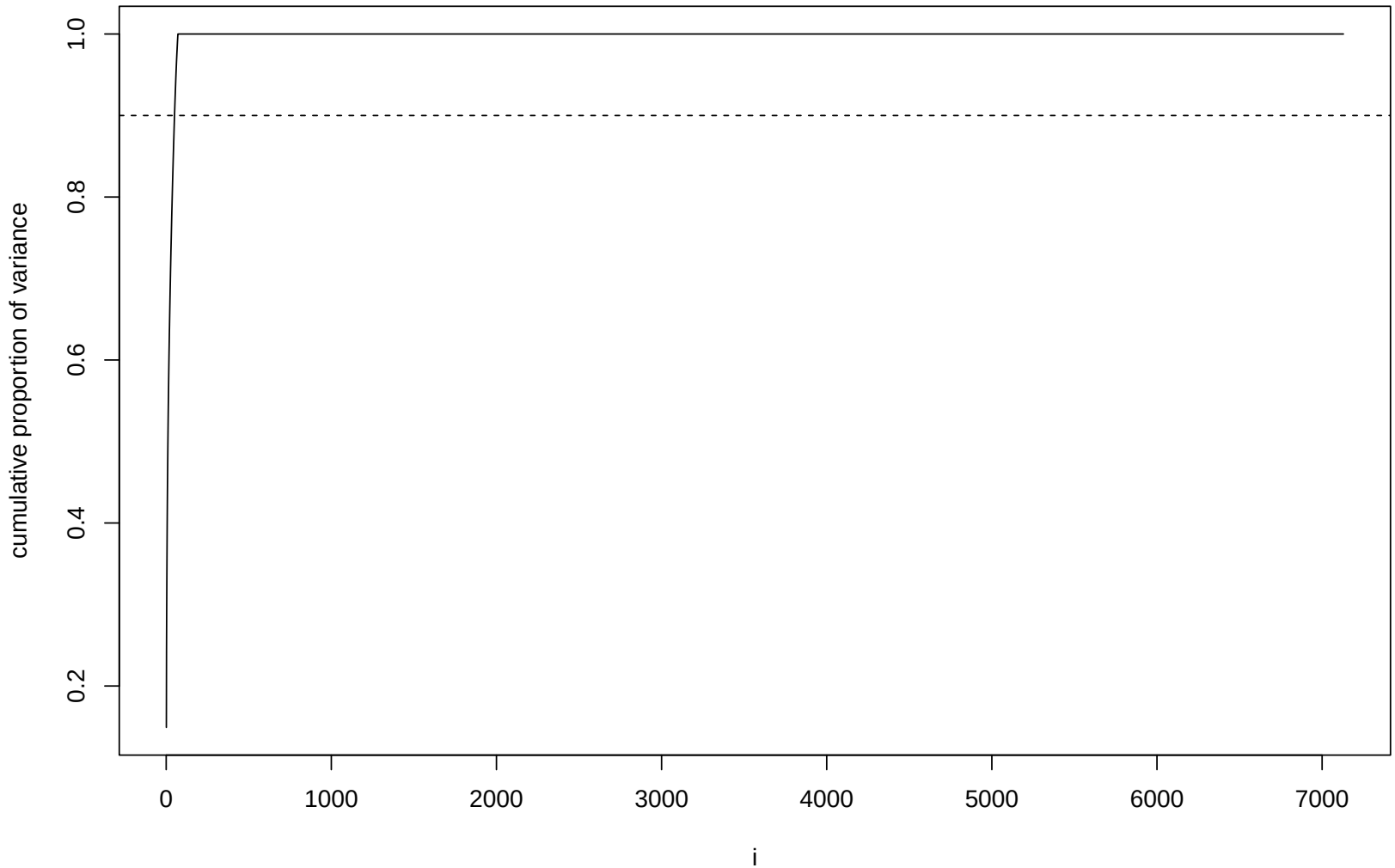
PCA

Since this is a problem of "$p >> n$", dimension reduction should be done before we implement a logistic regression model.

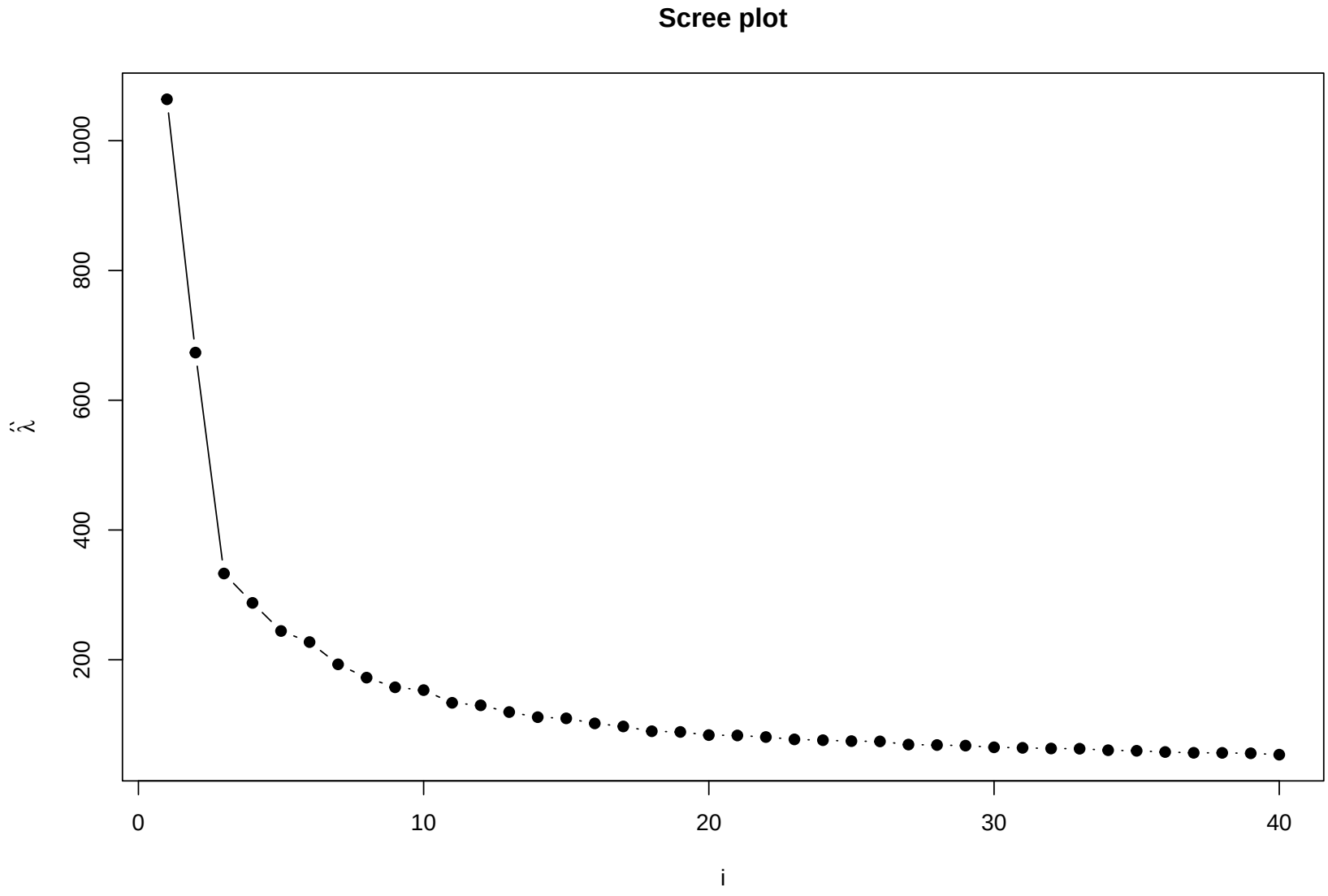
- First we standardized the gene expression level as their ranges vary a lot.
- Then we calculated the covariance matrix, eigen vectors, and cumulative variance as learned in class.

Cumulative Variance

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



Scree Plot



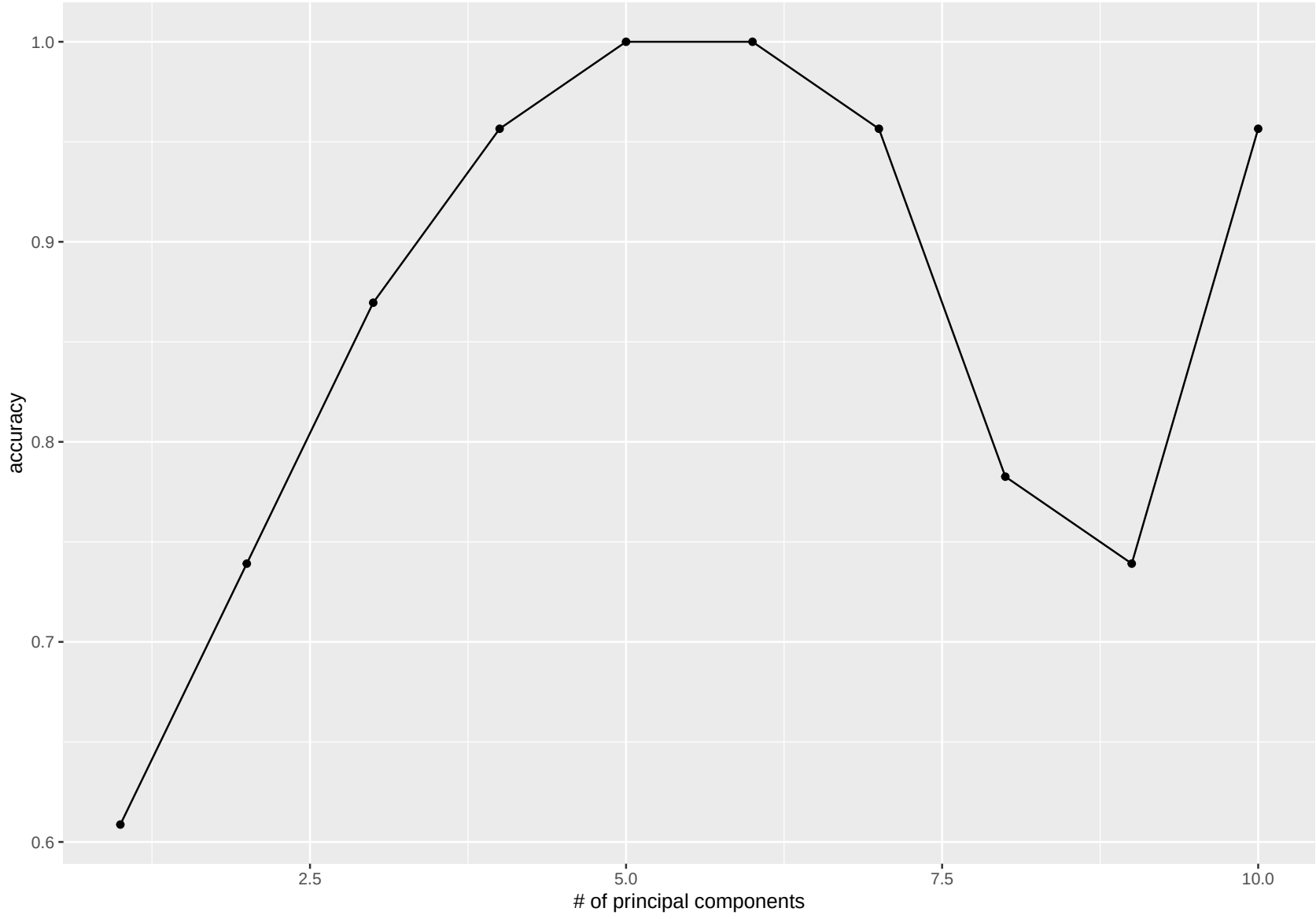
Scree Plot

But, of course, we cannot include all of the 70 components to the data as our sample size is only 72. Including all of them will cause overfit.

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.

Accuracy Plot



Accuracy

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.

Bootstrap

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.

Bootstrap Results

	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

Bootstrap result

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 bootstrap confidence interval show it is negatively correlated. So we will further explore the second component.

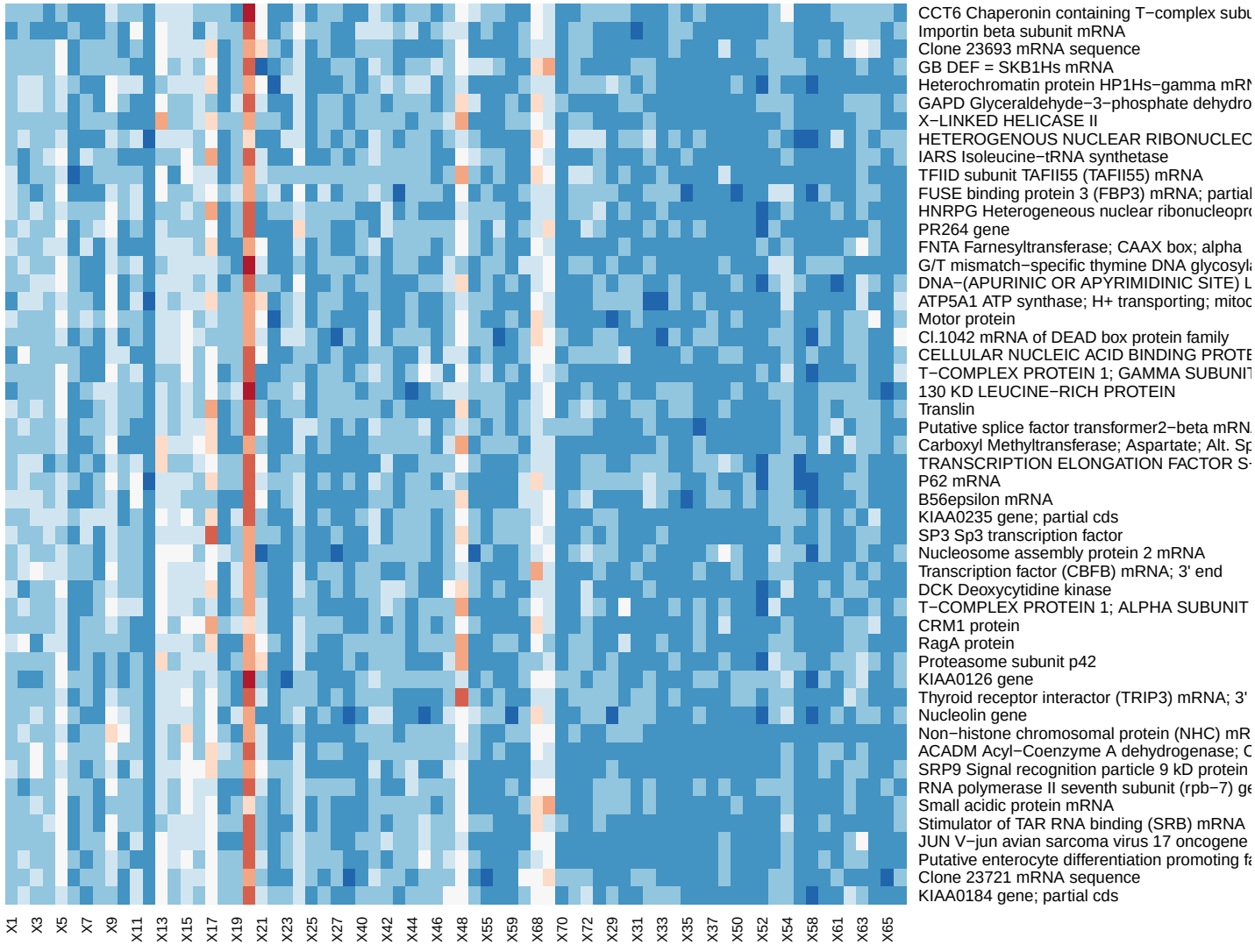
We want to look into it's marginal correlations to determine which gene expression contributes most to this component by calculating the marginal correlations.

Marginal correlation

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.

Heatmap of gene expression level



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.