Hepatitis C and Unsupervised Learning

Hanan Salim

Contents

1 Introduction

Hepatitis C is a viral infection caused by a blood borne pathogen, the hepatitis C virus (HCV). If left untreated this can lead to serious damage to the liver or cancer.

The virus attacks the liver which leads to an immune response, releasing various types of fibrosis proteins, like collagen, to repair the damage. Unfortunately, these proteins can build up within the liver and cause scarring. This scar tissue build up is called fibrosis. Over time this can lead to the death of liver cells and eventually liver failure.

To measure the progression of the disease, we can measure fibrosis. The stages are as follows: Stage 0: no fibrosis Stage 1: mild fibrosis without walls of scarring Stage 2: mild to moderate fibrosis with walls of scarring Stage 3: bridging fibrosis or scarring that has spread to different parts of the liver but no cirrhosis Stage 4: severe scarring, or cirrhosis

Our data has the following columns:

1. `X`: patient ID
2. `Category`: response variable
3. `Age`: age of the patient
4. `Sex`: sex of the patient
5. `ALB`: Albumin
6. `ALP`: Alkaline phosphatase
7. `ALT`: alanine amino-transferase

8. `AST`: aspartate amino-transferase
9. `BIL`: bilirubin
10. `CHE`: choline esterase
11. `CHOL`: cholesterol
12. `CREA`: creatinine
13. `GGT`: γ-glutamyl-transferase
14. `PROT`: metalloproteinase 1

Our response variable has five categories:

1. Blood Donor
2. Suspect Blood Donor
3. Hepatitis
4. Fibrosis
5. Cirrhosis

Our goal in this project is to perform PCA and clustering on our data. We would also like to detect outliers via local outlier factor (LOF) by creating a binary variable from our response variable.

2 EDA and Feature Engineering

From our summary statement below we can see that we have missing values in many of our columns. We can see what percentage of our observations are missing in the table below. We handle these missing values, my using mean imputation. As the first figure shows, the imputed values for ALP show a similar distribution to the original data.

```
       X           Category              Age            Sex
 Min.   :  1.0   Length:615         Min.   :19.00   Length:615
 1st Qu.:154.5   Class :character   1st Qu.:39.00   Class :character
 Median :308.0   Mode  :character   Median :47.00   Mode  :character
 Mean   :308.0                      Mean   :47.41
 3rd Qu.:461.5                      3rd Qu.:54.00
 Max.   :615.0                      Max.   :77.00

      ALB             ALP              ALT              AST
 Min.   :14.90   Min.   : 11.30   Min.   :  0.90   Min.   : 10.60
 1st Qu.:38.80   1st Qu.: 52.50   1st Qu.: 16.40   1st Qu.: 21.60
 Median :41.95   Median : 66.20   Median : 23.00   Median : 25.90
 Mean   :41.62   Mean   : 68.28   Mean   : 28.45   Mean   : 34.79
 3rd Qu.:45.20   3rd Qu.: 80.10   3rd Qu.: 33.08   3rd Qu.: 32.90
 Max.   :82.20   Max.   :416.60   Max.   :325.30   Max.   :324.00
 NA's   :1       NA's   :18       NA's   :1
      BIL             CHE              CHOL             CREA
 Min.   :  0.8   Min.   : 1.420   Min.   :1.430   Min.   :   8.00
 1st Qu.:  5.3   1st Qu.: 6.935   1st Qu.:4.610   1st Qu.:  67.00
 Median :  7.3   Median : 8.260   Median :5.300   Median :  77.00
 Mean   : 11.4   Mean   : 8.197   Mean   :5.368   Mean   :  81.29
 3rd Qu.: 11.2   3rd Qu.: 9.590   3rd Qu.:6.060   3rd Qu.:  88.00
 Max.   :254.0   Max.   :16.410   Max.   :9.670   Max.   :1079.10
                                  NA's   :10
```

```
      GGT              PROT
 Min.   :  4.50   Min.   :44.80
 1st Qu.: 15.70   1st Qu.:69.30
 Median : 23.30   Median :72.20
 Mean   : 39.53   Mean   :72.04
 3rd Qu.: 40.20   3rd Qu.:75.40
 Max.   :650.90   Max.   :90.00
                  NA's   :1
```

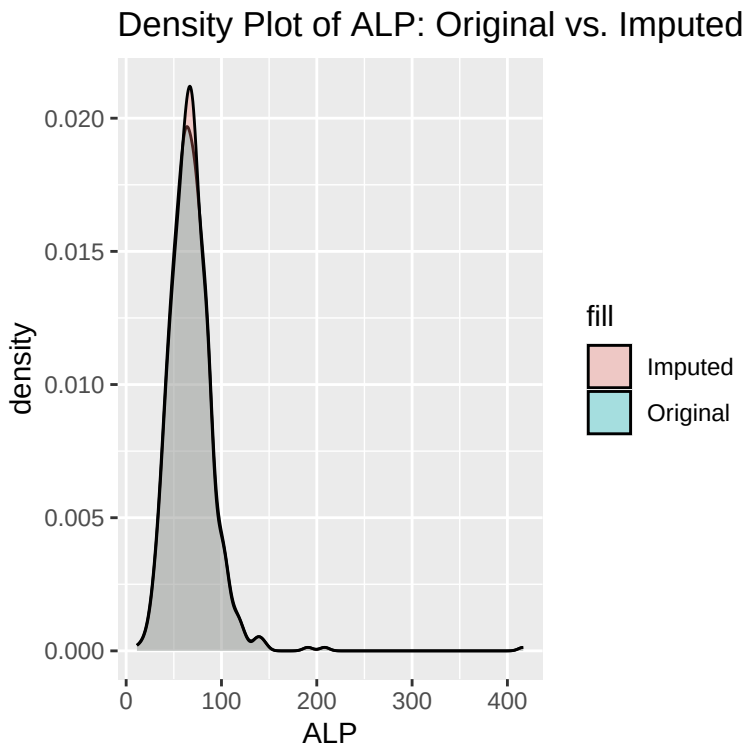


Figure 1: Imputed vs Original Data

We would now like to check the distribution of our features. From the second figure, we can see that our data is fairly skewed. To fix the skew, we would typically apply a log transformation but since our data has negative values and zeros, we instead apply a negative log transformation shown below.

$$T(x) = \text{sign}(x) \cdot \log(|x| + 1)$$

From the log transformed plot, we see that our data is less skewed but also shows better seperation between each category.

3 PCA and Clustering

From our scree plot, we can see that after 5 components,we see a leveling off. From the table, we can see that the first component captures about 25% percent of the variance and the first 5 components, all together, capture about 75%.

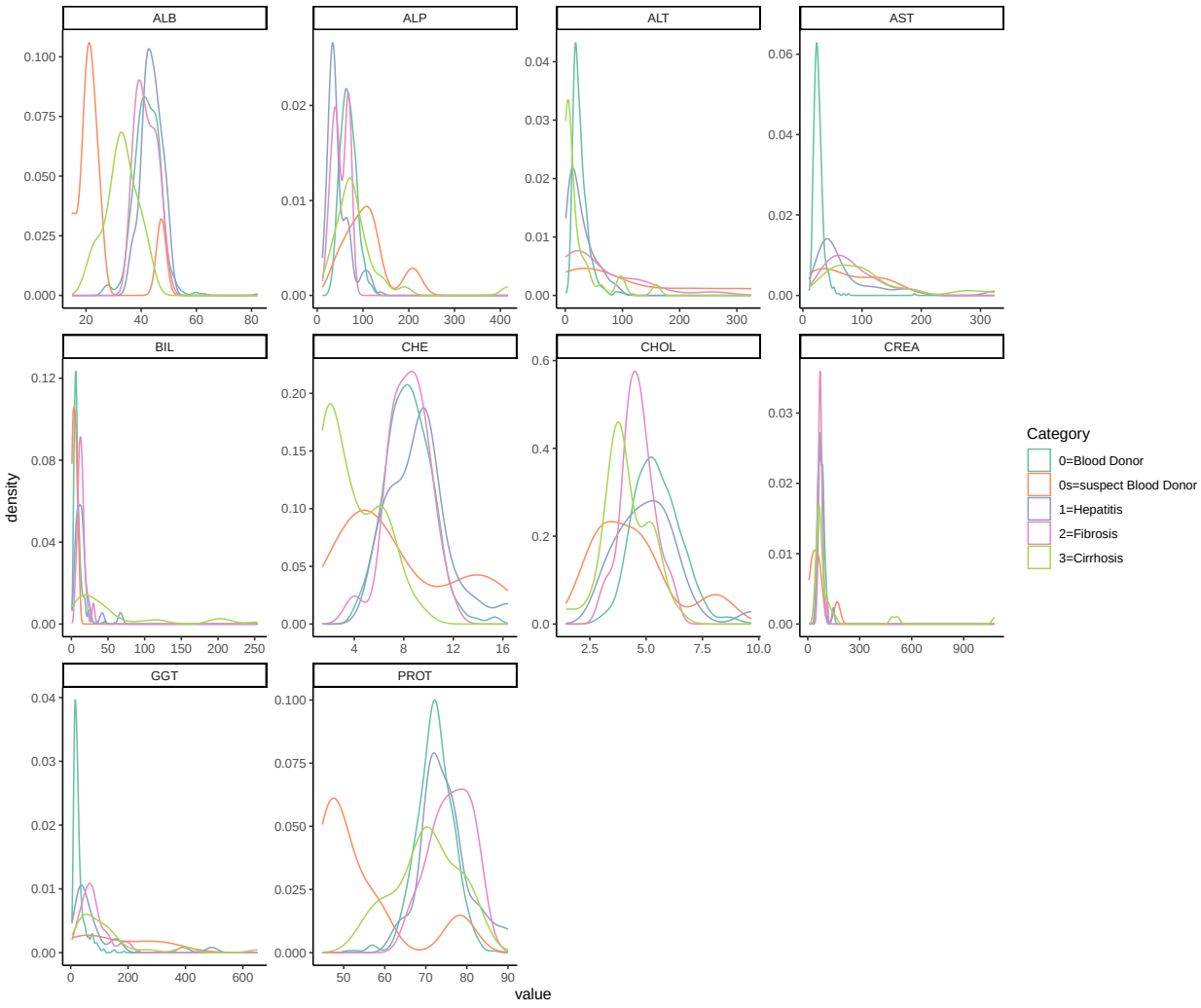


Figure 2: Density plots of features across disease progression

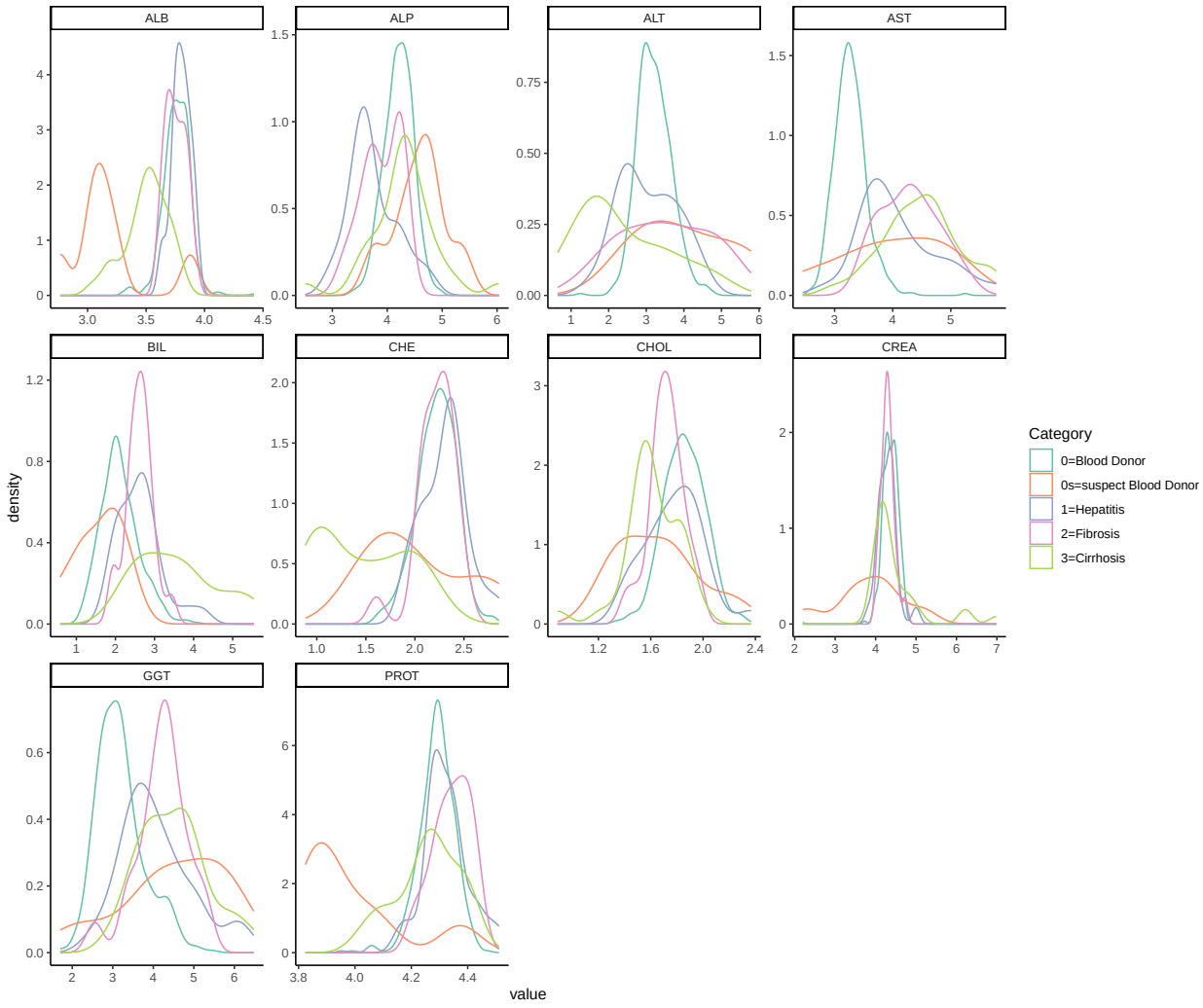


Figure 3: Normalized density plots of features across disease progession

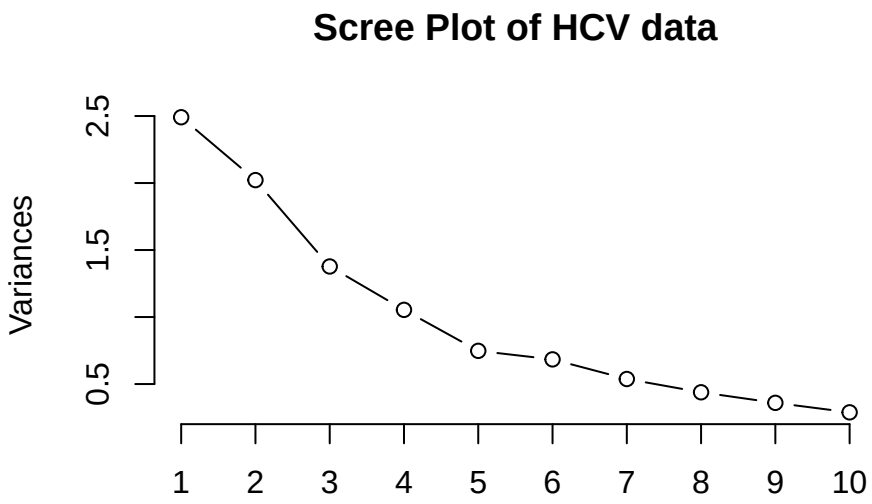


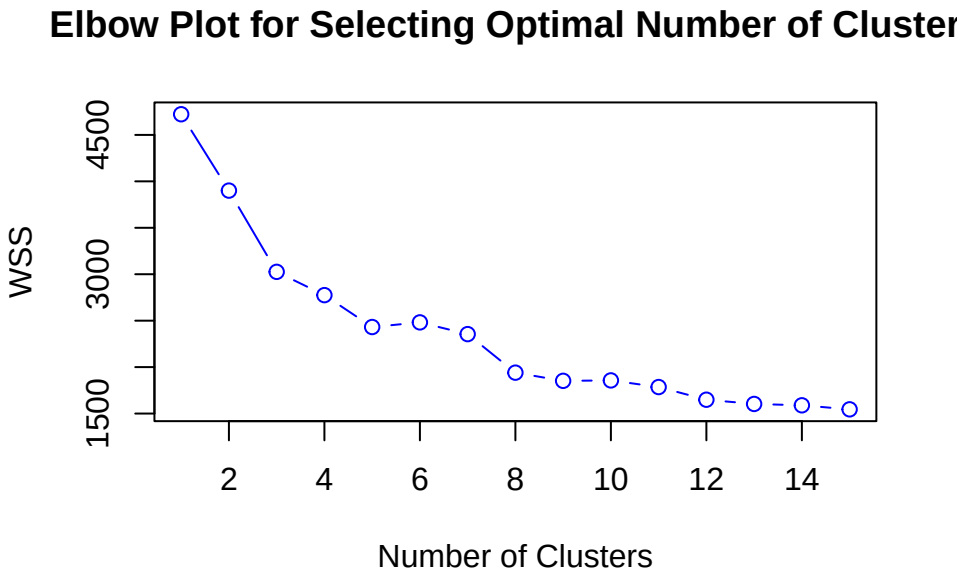
Table 1: The importance of each principal component

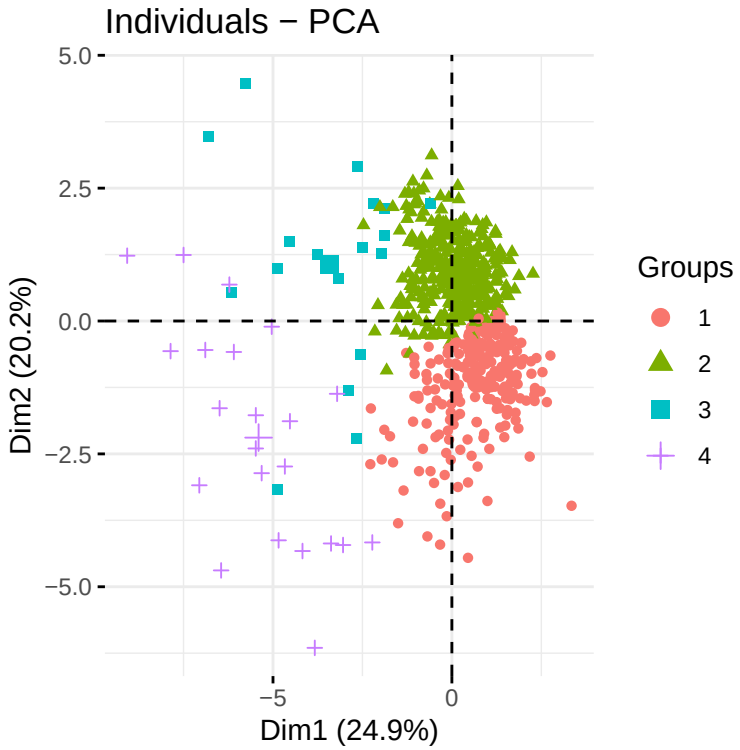
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.578271	1.421901	1.173598	1.026386	0.864736	0.827083	0.733231	0.662210	0.599771	0.537350
Proportion of Variance	0.249090	0.202180	0.137730	0.105350	0.074780	0.068410	0.053760	0.043850	0.035970	0.028870
Cumulative Proportion	0.249090	0.451270	0.589010	0.694350	0.769130	0.837540	0.891300	0.935150	0.971130	1.000000

4 Clustering Raw Data

We would like to use this data to perform k-means cluster. First, we would like to determine how many clusters are optimal. Intuitively, we know our data should have 4-5 clusters. To double check, we create an elbow plot which agrees with our initial intuition but the WSS score does not level off as sharply.

Furthermore, from our clustering plot, we can see that our clusters don't look great. This not surprising since PC1 and PC2 only capture about 45% of the variance.





5 Anomaly Detection

We begin by creating a binary variable from our response variable. We the `category` column was labeled as either Blood Donor or suspect Blood Donor, we label it as 0. Otherwise we label is as 1. From there we, implement LOF on our original data, using multiple k values ($k = 50, 100, 125, 150, 175$). From the output,

we can see that the larger values have a better area under the curve but with dimishing returns. Thus, in our case, we would choose $k = 175$ as our value.