Data Mining Project Report

Ana Mangarovska, Anastazija Kovachevikj, Guilherme Soares, Tiago Oliveira

April 18, 2023

1 Introduction and Objectives

Diffuse large-B-cell lymphoma is the most prevalent type of lymphoma in adults; nevertheless, the administration of anthracycline-based chemotherapy on patients that exhibit this disease yields a very low cure rate.

The study of microarray features has led to the identification of two subgroups of lymphomas, each one associated with different outcomes after chemotherapy. Moreover, these gene-expression profiles have also been used in survival analysis not only to further refine the previous distinction of lymphomas, but also to formulate a molecular predictor of the success of chemotherapy. Hence, gene-expression micro-array data has shown to be a promising tool in prognostics that deserves full attention.

In this project, we intend to analyse and apply some analysis techniques to a real dataset resulting from a study on this subject. The `patients.csv` file contains data from 240 patients (identified by their LYM number), divided into training set and validation set, data related to their follow-up in the study, data relating to genetic expression and the response to each of the 5 signatures tested. The `microarray.csv` file contains data relating to the expression of each of the 7292 genes in each of the patients.

Our objectives include a preliminary analysis of the data, data preprocessing, a check for missing values and impute them if exist, Outlier Detection, Principal Component Analysis, testing different clustering methods and performing a linear discriminant analysis.

Note that, while doing data preprocessing, we verified that there were some patients in the `microarray.csv` dataset that did not exist in the `patients.csv` dataset and, therefore, we only kept the data from `microarray.csv` corresponding to existing patients in the `patients.csv` table.

2 Missing Values Analysis

In the beginning of our study, we analysed the order of magnitude of the missing values present in our dataset. As we can see in Fig. 1, there are genes (variables) that have a significant number of missing values (some more than 100, that is, data related to those genes is missing in 100 or more samples, which is relevant, considering that our dataset only has 240 samples). To reduce the impact of the missing values in our posterior analysis, and because we have more than 7000 variables, we decided to remove the variables that presented more than 5% of missing values. We verified that there were 4992 genes with less than 10 % of missing values and 4035 genes with less than 5% of missing values. To deal with the missing values of those 4035 genes, we proceeded with some value imputation, that is, filling the missing entries with a specific value. To do that, we used the K Nearest Neighbors method [5]. This approach works by defining a number of neighbours K to consider and a strategy for the prediction of the value to be imputed. In our case, after testing several values of K (1, 2, 3, 5, 8, 10) and choosing the Euclidean distance to give greater influence to closer neighbors and less influence to those which are further away in the prediction of the values, we decided to use K = 5. Although the KNN method can be sensitive to outliers, it can be more accurate than the average, median or most frequent imputation methods.

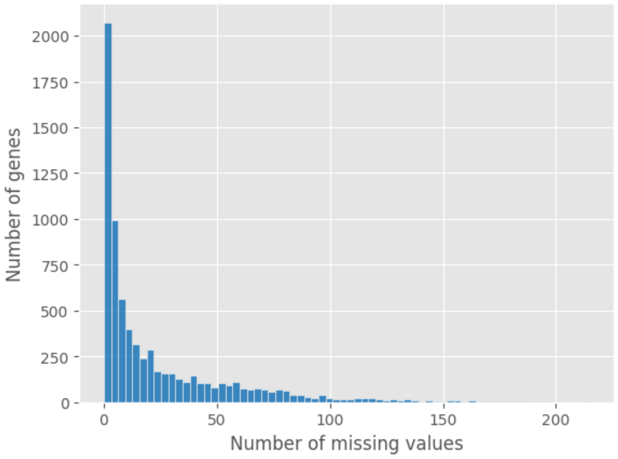
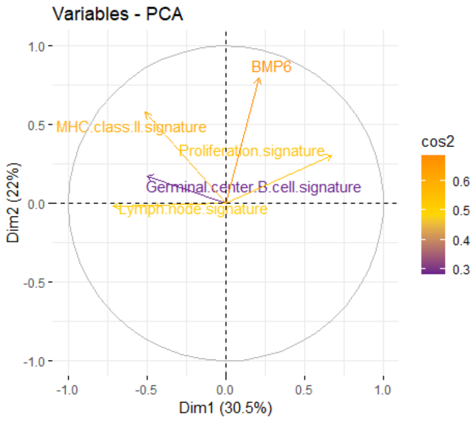


Figure 1: Number of missing values vs Number of genes

3 PCA

3.1 PCA on Signatures

Given the five signatures and its expression in all individuals we decided to apply Principal Component Analysis (PCA) [2], in order to understand which signatures are more relevant to the outcome (dead or alive) for the individuals. We plotted the contribution of each signature for the first two PCs in Fig.2. Then, we added the individuals projected in the same space - see Fig.3 - and coloured each dot according to each individual's outcome. Red indicates alive, while black means the individual passed away.



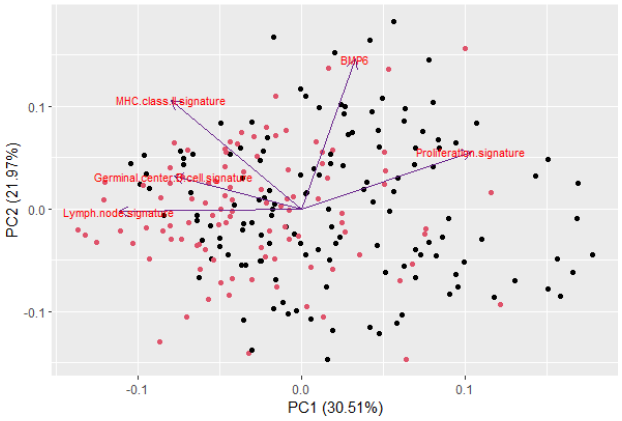


Figure 2: Contribution of each signature to the first two PCs. The five signatures were projected in two dimensions. The square of the cosine is captured by the coloured label.

Figure 3: Outcome of all individuals projected in the plane obtained using PC1 and PC2. The colour of each dot indicates if the outcome is alive (red) or dead (black). The five signatures were also projected.

By taking a closer look to these two plots, we observed several interesting aspects. First, we see that the majority of dead people (black dots) are on the right side of the graph of Fig.3, while the red dots (favorable outcome) are mostly on the left side. This agrees with the conclusions presented on the paper [4], as it states that the signatures *MHC class*, *Germinal center B-cell* and *Lymph node* yield a good outcome. Notice that the projections of these signatures point to the left as well. On the other hand, signatures correlated with a poor outcome in the paper - *Proliferation* and others - are pointing to the right side, just like the black dots.

Nevertheless, it is also worth noticing that not all the signatures are well represented by the PCA projections. Particularly, the *Germinal B-cell* signature has a very poor representation, since its cosine squared is quite close to zero - see Fig. 2.

3.2 PCA on Genes

3.2.1 Classical PCA

Regarding the genes provided by the `microarray.csv` file, we decided to apply PCA in order to reduce dimensionality. Initially, our dataset consists of about 4000 genes, which explains the need to reduce the dimensionality, since it is not practical to handle such a big number of variables. Note that, even though the initial dataset has over 7000 genes, we deleted the ones with many missing values which left us with the 4000 genes metioned.

Before proceeding with the PCA, we normalized the data. This step is crutial since variables in a larger scale would explain more variability - which would not be correct. We opted to use the min-max normalization [3], which scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one. After applying PCA, we chose to retain the number of PCs needed to explain 80% of all the variance. This number is 62, as we can see in Fig.5. This was advantageous for our analysis, since we will not have to deal anymore with 4000 variables - instead, from now onward we have just 62 variables to deal with.

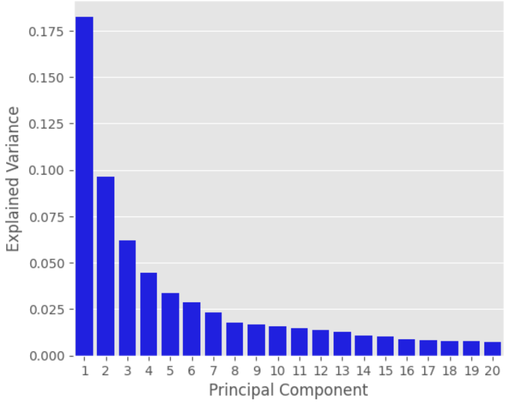


Figure 4: Variance explained by each Principal Component

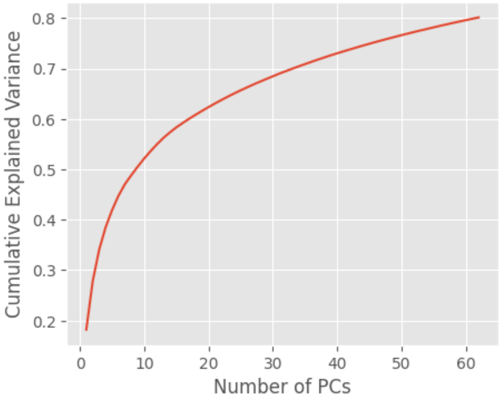


Figure 5: Cumulative variance in function of the number of Principal Components

3.2.2 Robust PCA

First of all, we shall observe that PCA is highly sensitive to outliers. This is due to the fact that PCA will rotate the coordinate axis in order to capture the directions of higher variance.

The presence of outliers in our dataset will precisely increase the variance in its direction since they will be far from the majority of the other points, by the definition of *outlier*. Thus, they will influence the directions of PCA in an unwanted but drastic way. Taking this into account it is clear that we need a PCA technique robust to outliers - and that is why we implemented Robust PCA [1]. In a few words, robust PCA will weight less points that are far away from the majority of points. This means they will have much less influence on the direction of the new axis.

Similarly to the classic PCA, we kept the first k components that can explain 80% of variability. This time k equals 57. This will be useful later, mostly when trying to cluster people into different sets in an attempt to recover the three groups identified in the paper. Other application of the robust PCA is the detection of outliers in the next section.

4 Next Steps

Normality Assessment;

Outlier Detection: Mahalanobis Distance and using PCA;

Clustering: Hierarchical Methods and Silhouette Analysis; Partitioning Methods and Silhouette Analysis;

Compare Clustering after performing the Classical PCA, a Robust PCA and without performing PCA;

Linear Discriminant Analysis;

Conclusions.

References

[1] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. Robpca: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.

[2] Richard Arnold Johnson and Dean W. Wichern. *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River, NJ, 5. ed edition, 2002.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] Andreas Rosenwald, George Wright, Wing C Chan, Joseph M Connors, Elias Campo, Richard I Fisher, Randy D Gascoyne, H Konrad Muller-Hermelink, Erlend B Smeland, Jena M Giltnane, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947, 2002.

[5] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.