Applying Compositional Data Analysis Methods To Complete Blood-Counts Data For Early COVID-19 Detection

Zhilong Zhang

A thesis
submitted in partial fulfillment of the
requirements for the degree of
Master of Science

University of Washington
2025

Committee:
Jan Graffelman
Ting Ye

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Applying Compositional Data Analysis Methods To Complete Blood-Counts Data For Early COVID-19 Detection

Zhilong Zhang

Chair of the Supervisory Committee:
Jan Graffelman
Biostatistics

The investigation of using complete blood-count (CBC) data analysis for COVID-19 infection diagnosis has been a topic of interest in the last couple of years. It could be used as an affordable complementary tool to RT-qPCR and is particularly useful for developing areas struggling to test their population or suffering from massive COVID-19 infection. However, previous research of using CBC data for COVID-19 infection classification didn't appreciate the compositional nature of white blood cell counts data. In this master's thesis, we treat white blood cell counts data as compositional variables and apply compositional data visualization methods, using biplots based on log-ratio principal component analysis. Also, we apply compositional classification models to detect COVID-19 infection, using log-ratio linear discriminant analysis. We successfully illustrate the efficacy of compositional methods by building a compositional classification model superior to traditional models and highlight the benefits of analyzing CBC data from a compositional perspective. In a database of symptomatic individuals, we achieve a classification rate of 85% for the PCR-test result using the main CBC composition with some additional blood characteristics.

Key words: COVID-19; CBC analysis; compositional data; log-ratio transform; multivariate analysis; principal component analysis; linear discriminant analysis; biplot; ROC curve

Acknowledgment

I would like to express my gratitude to Professor Jan Graffelman, who has been a wonderful instructor for this master's thesis. He provided insightful ideas that motivated this thesis and guided me with great patience when I met all kinds of challenges when composing it. Moreover, he has been an excellent teacher when I was studying relevant statistical methods for this thesis. When I had questions on technical details or the big picture, Jan Graffelman always answered them promptly and clearly. The whole process of composing this master's thesis has been a great pleasure and educational journey, and this wonderful experience is attributed to my thesis instructor, Jan Graffelman.

I would like to thank Professor Márcio Dorn, who was one of the collectors and maintainers of the dataset used in this thesis. He was very helpful when we had difficulty interpreting the dataset's irregular behavior. Moreover, he explained the low sensitivity issue of the PCR test in great detail, helping us address this problem in the discussion section. His expertise and timely reply are indispensable for us to complete this thesis.

I would like to give special thanks to Professor Kenneth Rice, Minh Vo, and Maggie Tarnawa for helping with logistical problems related to this master's thesis. They helped with format questions and credit registration problems.

Part of this master thesis has been used for writing a article, which will shortly be available as a preprint at `https://www.biorxiv.org/` and will submitted for publication in a scientific journal.

Contents

1 Introduction

The public health crisis of Coronavirus disease (COVID-19) is one of the most severe pandemics humans have experienced for the past few centuries [11], and it still causes large burdens to our society in many ways including economy, health services, life expectancy, etc. [12]. Nowadays, although we have RT-qPCR as the gold standard for diagnosing SARS-Cov2 infection [3], it remains impossible to employ this test massively in certain developing areas due to financial and administration problems. There exist other available diagnostic tests, such as antigen tests, which may be more accessible in areas with limited resources. On the other hand, in most scenarios those developing areas like Ecuador and Brazil need a well-functioning surveillance system more than others to avoid massive infection and control health care expenditure.

Motivated by this situation, we are trying to use the complete blood-count (CBC) data analysis as a cheap and efficient complementary tool to help us diagnosing potential COVID-19 infected patients [4]. In this classification problem, we have all of the common blood test results/indicators as our predictors and COVID-19 infection status (binary, given by RT-qPCR) as our outcome. The first step is to apply principal component analysis (PCA) [9] to the whole CBC dataset and five major white blood cells (WBC) counts data for preliminary data visualization and possible feature selection. Based on the prior knowledge, we know that five major types of white blood cell counts (see Figure 1 for the five major WBC counts: neutrophils, eosinophils, basophils, monocytes, and lymphocytes) will give us high diagnose/prediction power for many diseases as they are closely connected with immune response [15]. Moreover, they can be viewed as natural compositional data as each of the five WBC counts is part of a whole, the total WBC count. Therefore the second step is to apply linear discriminant analysis (LDA) on the raw WBC dataset and log-ratio transformed WBC compositional data. Finally, we compare the performance of LDA on the raw data and compositional data and discuss the advantage of combining compositional data analysis and traditional classification method like LDA in this COVID-19 detection scenario.

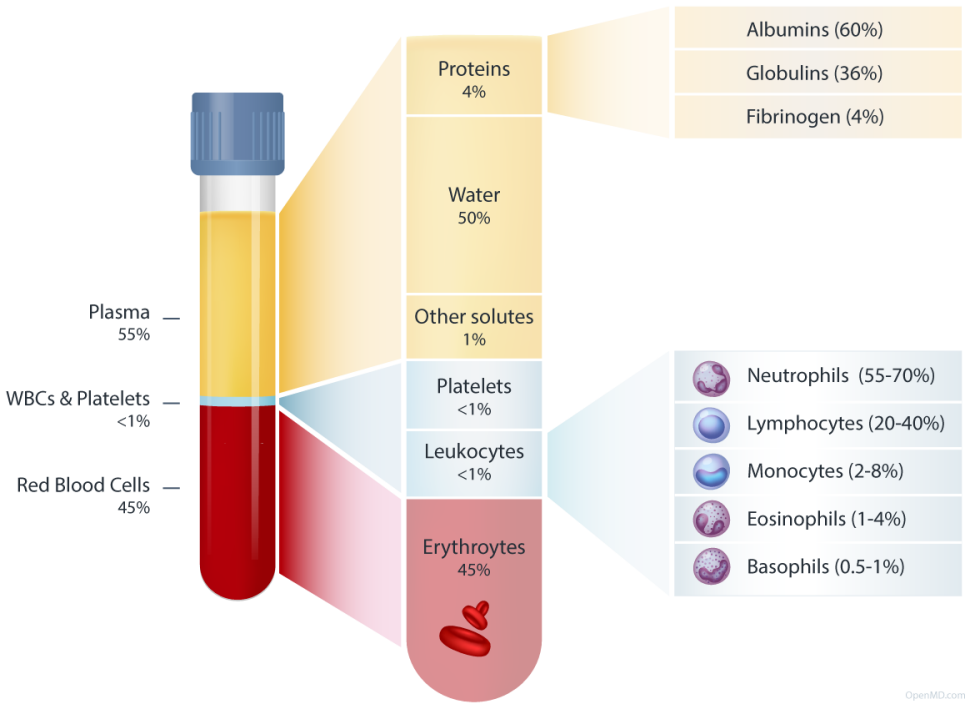

Figure 1: A demonstration of components of blood (source: `https://openmd.com/guide/blood-components`).

2 Statistical Methodology

In this section, we briefly introduce statistical methods used in this master thesis, including principal component analysis (PCA), linear discriminant analysis (LDA), the definition of compositional data and the log-ratio transform, log-ratio PCA (LR-PCA), and log-ratio LDA (LR-LDA).

2.1 PCA

Principal component analysis [9] is a linear dimension reduction technique that can be used for feature selection and multi-variate data visualization. The main motivation for PCA is to select a linear transformation of the variables, such that the first few transformed variables (called principal components) capture most of the variation of the data. Therefore, we can use the first few principal components (PCs) as the most important features that contain most information and ignore other components. If we only use the first two principal components, we are able to draw biplots giving the best two-dimensional approximation to the data matrix. In this thesis, we use PCA for the whole CBC dataset visualization and the WBC dataset visualization, and to identify patterns of COVID-19 infection status using the first two principal components.

To be more specific, let us present PCA in a sample based manner. Therefore, for the mean, covariance and variance symbol presented below, we are referring to the sample mean, sample covariance and sample variance respectively. Also, unless otherwise specified, we define all vectors as column vectors to avoid notation confusion. Assume we have n observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ that

are sampled independently from the same distribution (i.i.d. samples), and each has dimension p:

$$\mathbf{X}_{\text{raw}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \tag{1}$$

To make our expressions more compact, we denote

$$\mathbf{X} = \mathbf{X}_{\text{raw}} - \mathbf{1} \cdot \mathbf{m}^T \tag{2}$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones and $(\mathbf{m})_{p\times 1} = \frac{1}{n}\mathbf{X}^T \cdot \mathbf{1}$ is the mean vector with each entry equal to the mean of $\mathbf{X}_{\text{raw}}$ by columns. Now we have $\mathbf{X}$ as the centered data matrix, and we try to find a linear transformation from $\mathbf{X}$ to $\mathbf{F}$:

$$\begin{aligned} F_{i1} &= a_{11} \cdot x_{i1} + a_{12} \cdot x_{i2} + \ldots + a_{1p} \cdot x_{ip} \ , \\ F_{i2} &= a_{21} \cdot x_{i1} + a_{22} \cdot x_{i2} + \ldots + a_{2p} \cdot x_{ip} \ , \\ &\vdots \\ F_{ip} &= a_{p1} \cdot x_{i1} + a_{p2} \cdot x_{i2} + \ldots + a_{pp} \cdot x_{ip} \ , \end{aligned} \tag{3}$$

The equation above holds for any $i = 1, 2, \ldots, n$, with $\mathbf{F}_{n\times p} := (F_{ij})_{(n\times p)}$ and the transformation matrix $\mathbf{A}_{p\times p} := (a_{ij})_{p\times p}$ is chosen subject to:

- Columns of $\mathbf{F}$ (denote as $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_p$) are uncorrelated
- $\text{Var}(\mathbf{F}_1)$ is maximal among the choices of $\mathbf{A}$
- $\text{Var}(\mathbf{F}_1) \geq \text{Var}(\mathbf{F}_2) \geq \ldots \geq \text{Var}(\mathbf{F}_p)$
- $a_{i1}^2 + a_{i2}^2 + \ldots + a_{ip}^2 = 1$ for all $i = 1, 2, \ldots, p$.

According to Equation (3) and the constraint conditions, the normalized linear transformation $\mathbf{A}$ diagonalizes $Cov(\mathbf{X})$, and its diagonal entries are ordered from the biggest to the smallest. Based on the eigenvalue decomposition theory, the unique solution (regardless of direction) of the optimization problem above is given by:

$$\mathbf{A} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_p) \tag{4}$$

where $\mathbf{V}_r$ is the normalized (scaled to have norm 1) eigenvector corresponding to the r-th largest eigenvalue of the sample covariance matrix $\mathbf{S} := Cov(\mathbf{X})$. In other words, with the solution of this optimization problem, we have:

$$\mathbf{S} = \mathbf{A}\mathbf{D}_\lambda\mathbf{A}^T \tag{5}$$

with $\mathbf{D}_\lambda$ defined as the diagonal matrix with the eigenvalues of $\mathbf{S}$. Moreover, we can show that $\mathbf{D}_\lambda$ equals $Cov(\mathbf{F})$:

$$Cov(\mathbf{F}) = Cov(\mathbf{X} \cdot \mathbf{A}) = \mathbf{A}^T\mathbf{S}\mathbf{A} = \mathbf{A}^T\mathbf{A}\mathbf{D}_\lambda\mathbf{A}^T\mathbf{A} = \mathbf{D}_\lambda. \tag{6}$$

From Equation (6), we can see that the columns of $\mathbf{F}$ are uncorrelated variables and the first few columns explain most of the variance of our data. One thing to notice is that both data matrix $\mathbf{X}$

and transformed data matrix $\mathbf{F}$ have mean zero (for each variable/column) because of the centering operation applied beforehand.

Alternatively, the solution of a PCA analysis can be calculated from the singular value decomposition (SVD) framework. For consistency of notation, we still work at the sample level.

Define $\mathbf{X}_t := (1/\sqrt{n})\mathbf{X}$ and denote its SVD as

$$\mathbf{X}_t = \mathbf{U} \cdot \mathbf{D}_s \cdot \mathbf{A}^T \qquad (\mathbf{X}_t^T \cdot \mathbf{X}_t = \mathbf{S}). \tag{7}$$

where $\mathbf{U}$ is the $n \times r$ matrix of left singular vectors ($\mathbf{U}^T\mathbf{U} = I$), $\mathbf{D}_s$ is the $r \times r$ diagonal with non-increasing singular values, $\mathbf{A}$ is the $p \times r$ matrix right singular vectors ($\mathbf{A}^T\mathbf{A} = I$) and r is the rank of $\mathbf{X}_t$.

Now we have

$$\mathbf{F}_p = \mathbf{X} \cdot \mathbf{A} = \sqrt{n}\mathbf{U}\mathbf{D}_s\mathbf{A}^T\mathbf{A} = \sqrt{n}\mathbf{U}\mathbf{D}_s. \tag{8}$$

And $\mathbf{F}_p$ has uncorrelated columns because

$$Cov(\mathbf{F}_p) = \frac{1}{n}\mathbf{F}_p^T \cdot \mathbf{F}_p = \frac{1}{n}(\sqrt{n}\mathbf{U}\mathbf{D}_s)^T \cdot \sqrt{n}\mathbf{U}\mathbf{D}_s = \mathbf{D}_s^2 = \mathbf{D}_\lambda, \tag{9}$$

which is a diagonal matrix and its diagonal entries are ordered from the largest to the smallest.

Therefore, since the constraints of our optimization problem ask the unitary linear transformation $\mathbf{A}$ to diagonalize $\mathbf{S}$ and the diagonal entries are ordered from the largest to the smallest, we must pick the optimal linear transformation $\mathbf{A}$ provided by the SVD. This solution is unique (regardless of direction) as the SVD decomposition is unique.

Moreover, we can define $\mathbf{F}_s := \mathbf{F}_p\mathbf{D}_\lambda^{(-1/2)} = \sqrt{n}\mathbf{U}$ as the standardization of $\mathbf{F}_p$, then we have:

$$Cov(\mathbf{F}_s) = \frac{1}{n}\mathbf{F}_s^T \cdot \mathbf{F}_s = \mathbf{U}^T \cdot \mathbf{U} = \mathbf{I}_{p\times p}. \tag{10}$$

And we get two equivalent expressions of $\mathbf{X}$ to make two different kinds of biplots [5]:

$$\mathbf{X} = \sqrt{n}\mathbf{U}\mathbf{D}_s \cdot \mathbf{A}^T = \mathbf{F}_p\mathbf{A}^T = \mathbf{F}_s(\mathbf{A}\mathbf{D}_s)^T \tag{11}$$

By using $\mathbf{X} = \sqrt{n}\mathbf{U}\mathbf{D}_s \cdot \mathbf{A}^T = \mathbf{F}_p\mathbf{A}^T = \mathbf{F}_p\mathbf{G}_s^T$, we are able to draw a biplot called *form biplot*. In this way, we plot $\mathbf{F}_p$ as row coordinates and $\mathbf{G}_s = \mathbf{A}$ as the column coordinates (transformed coordinates of unit axis vector in $\mathbf{R}^p$). Selecting the first two columns of principal components $\mathbf{F}_p$ as data points and rows of first two eigenvectors (which are the first two column of $\mathbf{G}_s$) as arrows enable us to draw a two-dimensional plot. And it captures most of the information limited to two dimensions to visualize this multivariate dataset. Moreover, the form biplot preserves the Euclidean distance in the following sense:

$$\mathbf{X}(\mathbf{X})^T = n\mathbf{U}\mathbf{D}_\lambda\mathbf{U}^T = \mathbf{F}_p \cdot \mathbf{F}_p^T \ , \tag{12}$$

thus

$$d_E^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{f}_i - \mathbf{f}_j)^T(\mathbf{f}_i - \mathbf{f}_j) = d_E^2(\mathbf{f}_i, \mathbf{f}_j) \ . \tag{13}$$

Another common biplot can be obtained using $\mathbf{X} = \sqrt{n}\mathbf{U}\mathbf{D}_s \cdot \mathbf{A}^T = \mathbf{F}_s(\mathbf{A}\mathbf{D}_s)^T = \mathbf{F}_s\mathbf{G}_p^T$, and it is called *covariance biplot*. In this way, we plot $\mathbf{F}_s$ as the row coordinates and $\mathbf{G}_p = \mathbf{A}\mathbf{D}_s$ as the column coordinates. One property of $\mathbf{G}_p$ is that it equals the covariance between variables and components:

$$Cov(\mathbf{X}, \mathbf{F}_s) = \frac{1}{n}\mathbf{X}^T \cdot \mathbf{F}_s = \frac{1}{n}\sqrt{n}\mathbf{A}\mathbf{D}_s\mathbf{U}^T \cdot \mathbf{U}\sqrt{n} = \mathbf{A}\mathbf{D}_s = \mathbf{G}_p. \tag{14}$$

Therefore, selecting the first two columns of standardized principal components $\mathbf{F}_s$ as data points and the covariances of variables and first two standardized components (which is the first two column of $\mathbf{G}_p$) as arrows enable us to draw another two-dimensional plot. Moreover, Euclidean distances in the covariance biplot represent Mahalanobis distances of the original observations in the following sense:

$$\mathbf{X}\mathbf{S}^{-1}(\mathbf{X})^T = n\mathbf{U}\mathbf{U}^T = \mathbf{F}_s \cdot \mathbf{F}_s^T\ , \tag{15}$$

thus

$$d_M^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{f}_i - \mathbf{f}_j)^T(\mathbf{f}_i - \mathbf{f}_j) = d_E^2(\mathbf{f}_i, \mathbf{f}_j)\ . \tag{16}$$

All of the analysis and transformations above can be repeated if we replace the covariance matrix of our data with the correlation matrix of our data, and we can obtain the corresponding principal components and biplots (form biplot and covariance biplot) in the correlation sense. In practice, correlation-based PCA is used more often to deal with variables with incomparable units, and covariance-based PCA is used when all of our variables have the same unit. And the form biplot is used to visualize distance between observations, whereas the covariance biplot capitalizes on the correlation structure of the variables. In this thesis, we use correlation-based approach for the CBC dataset PCA, and using covariance-based approach for the WBC dataset PCA since WBC variables have the same units and similar scales.

Since the aim of the linear transformation $\mathbf{A}$ in PCA is to use transformed variables (principal components) to capture most of the data variance, a reasonable measurement of goodness-of-fit can be defined as the proportion of variance explained by the selected principal components. A rule of thumb is to choose the number of principal components that explains more than 80% of the total variance. The scree plot is a common way to show the variance of each principal component and to help us choose a reasonable number of PCs and visualize the proportion of variance explained by each PC. However, in practice there are situations that biplots of only the first two PCs are made that represent (far) less than 80% of the total variance, and they are still considered useful if they are interpretable. When explained variance is low, people need to be careful that important aspects of the data maybe missed and its full complexity not being captured.

2.2 LDA

Linear discriminant analysis [7][8] is a supervised classification technique used for datasets with categorical outcomes. It gives us a linear combination of variables that minimizes a combination of misclassification rate times the cost of misclassification. Moreover, it achieves dimension reduction by reducing p dimensional variables to $min(p, k-1)$ dimensional discriminators where k is the number of outcome levels. In our situation we have $k = 2$ as the outcome variable is either PCR-positive or PCR-negative; consequently, a single linear discriminant function is obtained.

Let us introduce two-group LDA in greater detail. Suppose we have $n = n_1 + n_2$ simple random samples drawn from the same population (i.i.d. samples), and we separate them in two groups.

We denote n_0 as the number of samples with outcome 0 (in our case, the patient has a PCR-negative condition), and n_1 as the number of samples with outcome 1 (in our case, the patient has a PCR-positive condition).

$$\mathbf{X}_{n_0} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_0 1} & x_{n_0 2} & \dots & x_{n_0 p} \end{bmatrix}, \qquad y \equiv 0 \tag{17}$$

$$\mathbf{X}_{n_1} = \begin{bmatrix} x_{(1+n_0)1} & x_{(1+n_0)2} & \dots & x_{(1+n_0)p} \\ x_{(2+n_0)1} & x_{(2+n_0)2} & \dots & x_{(2+n_0)p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(n_1+n_0)1} & x_{(n_1+n_0)2} & \dots & x_{(n_1+n_0)p} \end{bmatrix}, \qquad y \equiv 1 \tag{18}$$

We try to accurately classify any new observation $\mathbf{x} = (x_1, x_2, \dots, x_p)$ into $y = 0$ or $y = 1$, based on our current observations. In other words, we try to find a classification rule $\mathbf{R}$ that has a small cost of misclassification while we take prevalence into account.

We denote π_0 and π_1 as the population with outcome $y = 0$ and $y = 1$, and $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$ are their corresponding probability density functions with respect to predictor $\mathbf{x}$. One thing to notice is that a well-defined classification rule R should be a separation of the sample space $\mathbf{R} = \mathbf{R}_0 \sqcup \mathbf{R}_1$ (if $\mathbf{x}$ falls in $\mathbf{R}_0$, we classify it as π_0; if $\mathbf{x}$ falls in $\mathbf{R}_1$, we classify it as π_1). And we denote p_0 as the prior probability of π_0, p_1 as the prior probability of π_1. Based on the Bayes theorem, we can quantify the misclassification probability:

$$P(0|1) := P(\text{classified as 0 given true outcome is 1}) = P(\mathbf{x} \in R_0|\pi_1) = \int_{R_0} f_1(\mathbf{x}) d\mathbf{x} \tag{19}$$

$$P(1|0) := P(\text{classified as 1 given true outcome is 0}) = P(\mathbf{x} \in R_1|\pi_0) = \int_{R_1} f_0(\mathbf{x}) d\mathbf{x}. \tag{20}$$

We define $c(1|0)$ and $c(0|1)$ as the cost of misclassifying true negative as positive (false positive) and the cost of misclassifying true positive as negative (false negative) respectively. Using all the notations introduced above, we can define the expected cost of misclassification as the ultimate goal we try to minimize (and also as a measurement of performance of the classification technique):

$$ECM = c(0|1)P(0|1)p_1 + c(1|0)P(1|0)p_0. \tag{21}$$

In LDA, both f_1 and f_0 are chosen to be multivariate normal distributions with the same shape (same covariance matrix $\boldsymbol{\Sigma}$) and different means ($\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$):

$$\begin{aligned} f_1(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2} \cdot |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}_1)} \\ f_0(\mathbf{x}) &= \frac{1}{(2\pi)^{p/2} \cdot |\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}_0)} \end{aligned} \tag{22}$$

Since our choices of f_1 and f_0 have the same shape regardless of shifting, we are able to minimize ECM and obtain the optimal classification rule:

$$R_1 : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \geq \frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}; \qquad R_0 : \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} < \frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}. \tag{23}$$

And considering the specific expression of the multivariate normal distribution, the classification rule (23) is equivalent to :

$$\begin{aligned} R_1 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0) \geq ln\Big(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}\Big); \\ R_0 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0) < ln\Big(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}\Big). \end{aligned} \tag{24}$$

At the sample level, we plug in the sample mean vectors $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_0$ (both p dimensional) in the place of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$. Also, we plug in the pooled covariance matrix $\mathbf{S}_p = \frac{n_1-1}{n_1+n_0-2}\mathbf{S}_1 + \frac{n_0-1}{n_1+n_0-2}\mathbf{S}_0$ in the place of population covariance $\boldsymbol{\Sigma}$, where $\mathbf{S}_1$ and $\mathbf{S}_0$ are the within group sample covariance matrices. In this way, we are able to obtain the sample level classification rule (25) parallel with (24):

$$\begin{aligned} R_1 &: (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \mathbf{S}_p^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_0) \geq ln\Big(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}\Big); \\ R_0 &: (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \mathbf{S}_p^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)^T \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_0) < ln\Big(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}\Big). \end{aligned} \tag{25}$$

Notice that the left hand side of (25) is actually a linear function of predictor $\mathbf{x}$. If we define $\mathbf{a} := \mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$, we can write $y_i := \mathbf{a}^T\mathbf{x}_i$ and $\bar{y}_1 := \mathbf{a}^T\bar{\mathbf{x}}_1$; $\bar{y}_0 := \mathbf{a}^T\bar{\mathbf{x}}_0$. Then we write the ECM classification rule in a more compact way:

$$\begin{aligned} R_1 &: y_i \geq \frac{1}{2}(\bar{y}_1 + \bar{y}_0) + ln\Big(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}\Big); \\ R_0 &: y_i < \frac{1}{2}(\bar{y}_1 + \bar{y}_0) + ln\Big(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1}\Big). \end{aligned} \tag{26}$$

For the situation of equal misclassification cost, there is a more simple way to measure the performance of the LDA classification rate by computing the apparent error rate (APER, defined as the ratio of misclassification cases over total sample size). However, this ratio is not an unbiased estimation of the actual error rate $AER = P(0|1)p_1 + P(1|0)p_0$. To account for this bias, we can use the leave-one-out (LOO) method to train n models and make predictions for all individuals while avoid using the individual itself. Then we can have an unbiased error rate estimator using the LOO-based APER. The confusion matrix is another way to assess the LDA classification with equal misclassification cost. It is a matrix representing the summary statistics of right/wrong classification numbers in each outcome group. An example of a confusion matrix can be seen in the data analysis part (see Table 6).

Since LDA for two groups give us an one dimensional classifier based on the threshold $\frac{1}{2}(\bar{y}_1 + \bar{y}_0) + ln(\frac{c(1|0)}{c(0|1)} \cdot \frac{p_0}{p_1})$, we are able to vary this threshold to get different classification rules and their corresponding misclassification rate per group. For each classification rule, the true positive rate (TPR) is defined as the ratio of classified positive over true positive cases, and the false positive rate (FPR) is defined as the ratio of classified positive over true negative cases. With these two rates for each classifier, we can draw a receiver operating characteristic (ROC) curve [7] as another way to illustrate the performance of our LDA models. Moreover, we can use the area under the curve (AUC) to compare different LDA models. Notice that a 100% accurate classification rule will give us $AUC = 1$ while a truly random classification rule will give us $AUC = 0.5$ with a $y = x$ ROC curve. An example is shown is Figure 2, where the green line is the ideal ROC curve with

$AUC = 1$, the blue line is the random classification with $AUC = 0.5$, and the red line is an example ROC curve.

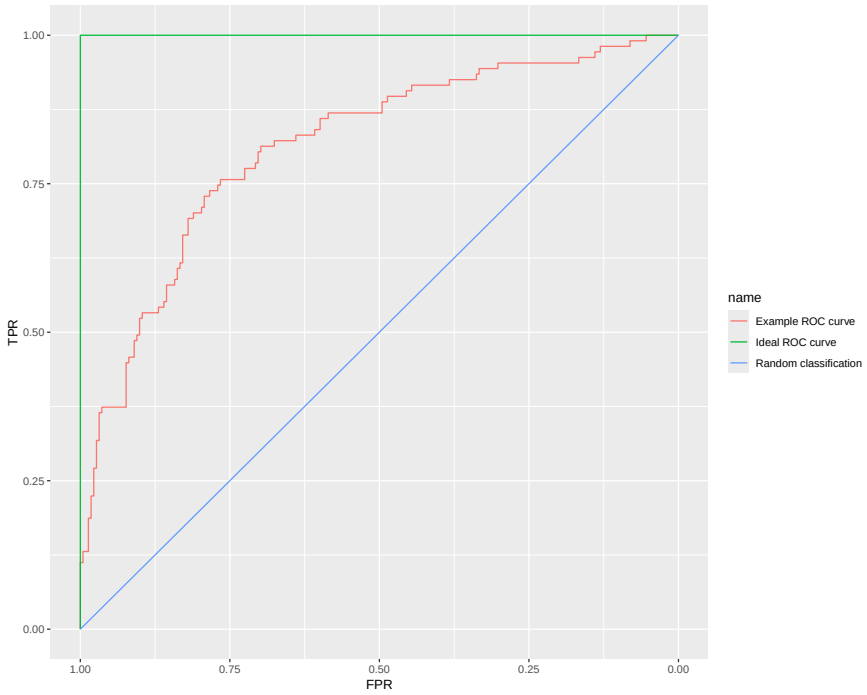

Figure 2: An example of ROC curves.

2.3 Compositional Data

Compositional data consists of variables that are parts of some whole. Proportions, percentages, and concentrations are common types of compositional data. When analysing those type of datasets, some issues come up that are hard to overcome by traditional statistical methods. For example, it is hard to define correlation and independence with respect to compositional data since all of them must sum up to a whole. This motivates us to build new methods and techniques to deal with this specific type of data, generally called methods for compositional data analysis (CoDA) [16].

To be more specific, let us define a *composition* of D parts to be a random vector:

$$\mathbf{x} = (x_1, x_2, \ldots, x_D). \tag{27}$$

And we define the domain of our compositional data $\mathbf{x}$ to be a *simplex* S^D:

$$S^D := \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D) | x_i > 0, \forall i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa \right\}. \tag{28}$$

Since we only care about the relative values of compositional data, we can re-express our data to a relative constant κ_0 (normally $\kappa_0 = 1$), and this operation is called *closure* operation $\mathcal{C}$:

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa_0 x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa_0 x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa_0 x_D}{\sum_{i=1}^{D} x_i} \right). \tag{29}$$

On the domain (S^D) where our compositional data lives, we need new a geometry instead of the Euclidean geometry in $\mathbb{R}^n$ due to the lack of linear vector space structure and corresponding metric. To equip S^D with a new vector space structure and the corresponding new geometry (called Aitchison geometry [1]), we define "addition" and "multiplication" in S^D, called perturbation and powering respectively:

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_D y_D] \in S^D, \quad \forall \mathbf{x}, \mathbf{y} \in S^D; \tag{30}$$

$$\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha] \in S^D, \quad \forall \mathbf{x} \in S^D \text{ and } \alpha \in \mathbb{R}. \tag{31}$$

It's easy to see that S^D with those two operations is an Abel group with external product, therefore a vector space. To measure distance and angle on S^D, we also need the inner product:

$$< \mathbf{x}, \mathbf{y} >_a := \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} ln\Big(\frac{x_i}{x_j}\Big) \cdot ln\Big(\frac{y_i}{y_j}\Big), \tag{32}$$

and we can define norm and distance induced by it.

Due to the nature of compositional data, three principles are proposed and should be fulfilled by any statistical method that we are trying to apply to compositional data [16]. They are *scale-invariance* (multiplying a composition by any positive value, or equivalently select any positive constant κ_0 shouldn't matter for the result of our analysis), *permutation invariance* (reording parts of the composition $\mathbf{x}$ should give us equivalent results in compositional data analysis), and *subcompositional coherence* (any analysis applied to the subcomposition should give us results coherent with the analysis applied to the full composition. For example, a well defined distance should satisfy that sub-composition distance is always less or equal than full composition distance).

Compositional data satisfy the three principles by first applying the *log-ratio transformation* to the compositional data and then apply classical statistical method to the transformed data. Three common log-ratio transformation are the additive log-ratio transformation [16] (alr):

$$alr(\mathbf{x}) = \Big[ln\Big(\frac{x_1}{x_D}\Big), ln\Big(\frac{x_2}{x_D}\Big), \ldots, ln\Big(\frac{x_{D-1}}{x_D}\Big)\Big]; \tag{33}$$

the centred log-ratio transformation [16] (clr):

$$clr(\mathbf{x}) = \Big[ln\Big(\frac{x_1}{g_m(\mathbf{x})}\Big), ln\Big(\frac{x_2}{g_m(\mathbf{x})}\Big), \ldots, ln\Big(\frac{x_D}{g_m(\mathbf{x})}\Big)\Big]; \tag{34}$$

where $g_m(\mathbf{x})$ is the geometric mean of the components of the composition $\mathbf{x}$:

$$g_m(\mathbf{x}) = \Big(\prod_{i=1}^{D} x_i\Big)^{(1/D)}, \tag{35}$$

and the isometric log-ratio transformation [16] (ilr):

$$ilr(\mathbf{x}) = clr(\mathbf{x})\mathbf{\Psi}^T. \tag{36}$$

In the definition of *ilr*, $\mathbf{\Psi}$ is a contrast matrix with dimension $(D-1)\times D$. And it satisfies:

$$\mathbf{\Psi}\cdot\mathbf{\Psi}^T=(\mathbf{I})_{(D-1)\times(D-1)},\qquad \mathbf{\Psi}^T\cdot\mathbf{\Psi}=(\mathbf{I})_{D\times D}-\frac{1}{D}\mathbf{1}_{D\times 1}\mathbf{1}^T. \tag{37}$$

Notice that all three log-ratio transformations have their advantages and disadvantages. For example, *ilr* is an isometry mapping between S^D and $\mathbb{R}^{D-1}$ but is computationally heavier, and its coordinates' expressions are based on the choice of a basis (or equivalently, the choice of $\mathbf{\Psi}$) thus they are not unique. Also, for *alr* one needs to specify the denominator in Equation (33) because our analysis should satisfy permutation-invariance thus the position of components doesn't matter.

2.4 LR-PCA

We introduce Log-Ratio PCA (LR-PCA) as the PCA method for compositional data. The Log-Ratio transformation part is chosen to be *clr* transformation (Equation (34)) although we can combine PCA with other log-ratio transformations as well. The advantage of choosing *clr* includes the uniqueness of our transformation (there is no need to choose a contrast matrix $\mathbf{\Psi}$, unlike *ilr* in Equation (36); or a denominator for *alr* in Equation (33)), preserving isometry between Aitchison distance of compositions and Euclidean distance of transformed coordinates (unlike *alr* in Equation (33)), etc. One thing to notice is that since every component in our composition naturally has the same unit, we prefer covariance-based PCA after the log-ratio transformation to a correlation-based PCA.

Continuing with notations introduced earlier, the coordinate of $clr(\mathbf{x})_i$ equals:

$$clr(\mathbf{x})_i=ln\Big(\frac{x_i}{g_m(\mathbf{x})}\Big)=ln\Big(\frac{x_i}{(\prod_{i=1}^D x_i)^{1/D}}\Big)=ln(x_i)-\frac{1}{D}\sum_{i=1}^D ln(x_i). \tag{38}$$

We note that the *clr* transformation amounts to a centering operation of the logarithmic scores. To write it in a matrix form, we have the log-transformed data $\mathbf{X}_l$:

$$\mathbf{X}_l=ln(\mathbf{X}); \tag{39}$$

and the *clr*-transformed data $\mathbf{X}_{clr}$:

$$\mathbf{X}_{clr}=\mathbf{X}_l\cdot\mathbf{H}_r; \tag{40}$$

and the column centered *clr* data $\mathbf{X}_{cclr}$:

$$\mathbf{X}_{cclr}=\mathbf{H}_c\cdot\mathbf{X}_{clr}=\mathbf{H}_c\cdot\mathbf{X}_l\cdot\mathbf{H}_r. \tag{41}$$

Where

$$\mathbf{H}_r:=\mathbf{I}_{D\times D}-\frac{1}{D}\mathbf{1}_{D\times 1}\cdot\mathbf{1}^T,\qquad \mathbf{H}_c:=\mathbf{I}_{n\times n}-\frac{1}{n}\mathbf{1}_{n\times 1}\cdot\mathbf{1}^T \tag{42}$$

are the centering matrix in the *clr* definition and before the PCA process respectively, and we call our double-centered log-transformed data $\mathbf{X}_{cclr}$. After the double-centering process, we can just apply standard PCA to $\mathbf{X}_{cclr}$, either by the spectral decomposition of its covariance matrix or by the SVD of $\mathbf{X}_{cclr}$.

Using the singular value decomposition, we can write

$$\frac{1}{\sqrt{n}}\mathbf{X}_{cclr} = \mathbf{UDV}, \tag{43}$$

then we are able to construct form biplots based on

$$\mathbf{F}_p = \frac{1}{\sqrt{n}}\mathbf{UD}, \qquad \mathbf{G}_s = \mathbf{V} \tag{44}$$

and covariance biplots based on

$$\mathbf{F}_s = \frac{1}{\sqrt{n}}\mathbf{U}, \qquad \mathbf{G}_p = \mathbf{VD}. \tag{45}$$

There are some special properties of a compositional biplot obtained by LR-PCA [2] that deserve to be mentioned. First is the isometry of *clr* and the distance preserving property of the form biplot (Equation (13)) ensures that the Euclidean distance between LR-PCA scores equals the Aitchison distance between data points:

$$d_{\text{Aitchison}}(\mathbf{x}_1, \mathbf{x}_2) = d_E(clr(\mathbf{x}_1), clr(\mathbf{x}_2)) = d_E(\mathbf{f}_1, \mathbf{f}_2), \tag{46}$$

the first equation can be easily checked using the definition of the Aitchison distance and the *clr* transformation.

Also, it is easy to see from the centering process that the origin in both biplots represents the geometric mean of the compositions. And the links between LR-PCA vectors represent pair-wise log-ratios: if x_1 and x_2 are the first and the second part of a compositional variable, then we have

$$clr(x_1) - clr(x_2) = ln\Big(\frac{x_1}{g_m(\mathbf{x})}\Big) - ln\Big(\frac{x_2}{g_m(\mathbf{x})}\Big) = ln\Big(\frac{x_1}{x_2}\Big). \tag{47}$$

The length of links between the vectors represents the standard deviation of the corresponding log ratio:

$$\begin{aligned} ||\mathbf{g}_i - \mathbf{g}_j||^2 &= \mathbf{g}_i^T\mathbf{g}_i + \mathbf{g}_j^T\mathbf{g}_j - 2\mathbf{g}_i^T\mathbf{g}_j \\ &= Var(clr(x_i)) + Var(clr(x_j)) - 2Cov(clr(x_i), clr(x_j)) \\ &= Var(clr(x_i) - clr(x_j)) \\ &= Var\Big(ln\Big(\frac{x_1}{x_2}\Big)\Big). \end{aligned} \tag{48}$$

Based on Equation (47) and Equation (48), we have the following observations for a compositional biplot obtained by LR-PCA: close to coincident biplot vectors suggest proportionality of parts; cosines of angles between links represent correlations between log-ratios; and collinear biplot vectors suggest a one-dimensional pattern for a subcomposition.

2.5 LR-LDA

For a compositional dataset, we can not directly apply LDA method for classification because of the structural singularity of covariance matrix $\mathbf{\Sigma}$ in Equation (22). However, we can use a generalized inverse (e.g. the Moore–Penrose inverse) to overcome this problem ([17]). Also, in order to make our method applicable to multi-class cases instead of two group dataset, we introduce log-ratio linear discriminant analysis (LR-LDA) in the framework of Fisher's multi-class discriminant analysis.

To be more specific, let us introduce multi-class Fisher's discriminant analysis first. Fisher's discriminant analysis aims to find a linear transformation $\mathbf{a}$ of our random vector $\mathbf{x}$ that maximizes the ratio of variability between groups to variability within groups.

$$\underset{\mathbf{a}}{\text{argmax}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}, \tag{49}$$

where $\mathbf{B}$ is defined as the variability between k groups and $\mathbf{W}$ is defined as the variability within k groups:

$$\mathbf{B} := \sum_{i=1}^{k} n_i(\overline{\mathbf{x}_i} - \overline{\mathbf{x}})(\overline{\mathbf{x}_i} - \overline{\mathbf{x}})^T \tag{50}$$

$$\mathbf{W} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}_i})(\mathbf{x}_{ij} - \overline{\mathbf{x}_i})^T. \tag{51}$$

We denote n_i as the sample size of group i and n as the total sample size of whole dataset. Also, $\overline{\mathbf{x}_i}$ is the mean vector of group i and $\overline{\mathbf{x}}$ is the total mean vector of the whole dataset. $\mathbf{x}_{ij}$ is the data vector of case j that belongs to the group i. Notice that we have the following decomposition of the total variability, expressed by the total sum of squares:

$$\mathbf{T} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})^T = \mathbf{B} + \mathbf{W}. \tag{52}$$

By the eigenvalue decomposition theory, we have the solution to the objective function (49) above as

$$\mathbf{W}^+\mathbf{B}\mathbf{a} = \lambda\mathbf{a}; \tag{53}$$

or written in a matrix form,

$$\mathbf{W}^+\mathbf{B}\mathbf{A} = \mathbf{A}\mathbf{D}_\lambda, \tag{54}$$

where columns of $\mathbf{A}$ are eigenvectors satisfying $\mathbf{A}^T\mathbf{S}_p\mathbf{A} = \mathbf{I}$ with $\mathbf{S}_p$ the pooled covariance matrix defined before. Notice that the dimension of solution is the rank of matrix $rank(\mathbf{W}^+\mathbf{B}) = min(D-1, k-1)$ as our compositional variables are constrained and have $dim = D-1$. For the compositional dataset, we use $\mathbf{W}^+$ to denote $\mathbf{W}$'s Moore-Penrose pseudo-inverse since $\mathbf{W}$ is singular. If $\mathbf{W}$ has an inverse, then the generalized inverse $\mathbf{W}^+$ just equals the normal matrix inverse: $\mathbf{W}^+ = \mathbf{W}^{-1}$.

Equivalently, we can solve the objective function by the singular value decomposition and it gives us the same solution of objective function (49). For the ease of notation, let us assume we have centered the dataset (this is automatically true after the double-centered log-ratio transformation as in Equation (41)). Denote the diagonal matrix $\mathbf{D}_w$ as $\mathbf{D}_w := \mathbf{Diag}(\text{proportion of group } i)$ and group mean matrix $\mathbf{M}_c^T$ as $\mathbf{M}_c^T := \left(\overline{\mathbf{x}_1} - \overline{\mathbf{x}},\ \overline{\mathbf{x}_2} - \overline{\mathbf{x}}, \ldots,\ \overline{\mathbf{x}_k} - \overline{\mathbf{x}}\right)$. Then the singular value decomposition has the following expression:

$$\mathbf{K} := \mathbf{D}_w^{1/2}\mathbf{M}_c(\mathbf{S}_p^+)^{1/2} = \mathbf{U}\mathbf{D}_\lambda^{1/2}\mathbf{V}. \tag{55}$$

Notice that $\mathbf{S}_p$ is just a reweighed version of $\mathbf{W}$, as

$$\mathbf{S}_p := \sum_{i=1}^{k} \frac{n_i - 1}{n - k} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}_i})(\mathbf{x}_{ij} - \overline{\mathbf{x}_i})^T. \tag{56}$$

Then we can see that the SVD expression is actually equivalent with the eigenvalue decomposition expression. First notice that:

$$\begin{aligned} \mathbf{D}_n := \{Diag(n_1, \ldots, n_k)\}^{1/2} &= n^{\frac{1}{2}} \mathbf{D}_w \\ \mathbf{D}_n \mathbf{M}_c^T \mathbf{M}_c \mathbf{D}_n &= \mathbf{B} \\ (\mathbf{S}_p^+)^{1/2} \mathbf{D}_n^{-1} &\simeq (\mathbf{W}^+)^{1/2} \end{aligned} \tag{57}$$

Then we have:

$$\begin{aligned} \mathbf{K}^T \mathbf{K} &= \Big((\mathbf{S}_p^+)^{1/2} \mathbf{M}_c^T \mathbf{D}_w^{1/2}\Big)\Big(\mathbf{D}_w^{1/2} \mathbf{M}_c (\mathbf{S}_p^+)^{1/2}\Big) \\ &= \Big((\mathbf{S}_p^+)^{1/2} \mathbf{D}_n^{-1} \mathbf{D}_n \mathbf{M}_c^T\Big) \mathbf{D}_w \Big(\mathbf{M}_c \mathbf{D}_n \mathbf{D}_n^{-1} (\mathbf{S}_p^+)^{1/2}\Big) \\ &\simeq \Big(\frac{1}{n} (\mathbf{W}^+)^{1/2} \mathbf{D}_n \mathbf{M}_c^T\Big) \mathbf{D}_w \Big(\mathbf{M}_c \mathbf{D}_n (\mathbf{W}^+)^{1/2}\Big); \end{aligned} \tag{58}$$

and

$$\mathbf{W}^+ = (\mathbf{W}^+)^{1/2} (\mathbf{W}^+)^{1/2} \tag{59}$$

$$\mathbf{B} = \Big(\mathbf{D}_n \mathbf{M}_c^T\Big)\Big(\mathbf{M}_c \mathbf{D}_n\Big). \tag{60}$$

In equation (57) and (58), the last equations are only approximately equal in order to compensate for the difference between $n_i - 1$ and n_i, and multiplying a constant n^{-1} won't affect the result of the SVD or eigenvalue decomposition. And $\mathbf{D}_w$ is the prevalence of the different groups. This term was ignored in the Fisher's objective function (49) as it was assumed to be an identity matrix. Therefore, if we ignore n^{-1} and $\mathbf{D}_w$, we have built the equivalence between the SVD expression (55 and 58) and eigenvalue decomposition expression (49).

The optimal transformation for the SVD is given by $(\mathbf{S}_p^+)^{1/2}\mathbf{V}$ and it equals to the eigen-vector matrix $\mathbf{A}$ obtained in eigen decomposition Equation (54):

$$\mathbf{V}^T \mathbf{K}^T \mathbf{K} \mathbf{V} \Leftrightarrow \mathbf{A}^T \mathbf{W}^+ \mathbf{B} \mathbf{A} \qquad \text{with the constraint: } \mathbf{A}^T \mathbf{S}_p \mathbf{A} = \mathbf{I}. \tag{61}$$

$$\text{as we have: } (\mathbf{S}_p^+)^{1/2} \mathbf{V} = \mathbf{A} \tag{62}$$

The SVD method is attractive since it provides a way to draw biplots, as now we have the row coordinates of the group means and the column coordinates of transformed vectors:

$$\mathbf{F}_p = \mathbf{D}_w^{-1/2} \cdot \mathbf{U} \cdot \mathbf{D}_\lambda^{1/2}, \tag{63}$$

$$\mathbf{G}_s = \mathbf{V}. \tag{64}$$

Now we can project our compositional dataset $\mathbf{X}_{cclr}$ introduced in Equation (41) to the optimal linear transformation and obtain the scores of the individual observations for LR-LDA classification in $min(D - 1, k - 1)$ dimensions:

$$\mathbf{F} := \mathbf{X}_{cclr} \cdot ((\mathbf{S}_p^+)^{1/2} \mathbf{V}). \tag{65}$$

With the classification scores of our dataset, we can compare them with the group mean coordinates and decide their predicted group based on the Euclidean distance from the group mean in the discriminant space. To be more specific, we have the Euclidean distance between score of the individual observation $\mathbf{F}_j$ (j-th row of $\mathbf{F}$, $j = 1, 2, \ldots, n$) and the score of i-th group mean $\mathbf{F}_{p_i}$ (i-th row of $\mathbf{F}_p$, $i = 1, 2, \ldots, D$):

$$d_{ij} = \sqrt{(\mathbf{F}_j - \mathbf{F}_{p_i})^T(\mathbf{F}_j - \mathbf{F}_{p_i})} \tag{66}$$

The predicted group of case j is the one that minimizes the distance over all i: $\underset{i}{\operatorname{argmin}}\{d_{1j}, d_{2j}, \ldots, d_{Dj}\}$. Alternatively, classification can be based on the posterior probabilities, which are derived from the linear classifier assuming a normal distribution with the same covariance matrix but different means for the two groups. Notice that Fisher's discriminant analysis gives us the same classification result as the LDA method part if we assume identical covariance matrices in the two group LDA and don't allow singularity problem to happen, as their final solution of classifier thresholds are equivalent.

3 Database and Data Cleaning

In this section, we introduce our dataset and its variables, with some summary statistics. Moreover, we clean our dataset to exclude abnormal medical cases.

3.1 Variable description

Our original CBC dataset was obtained between 2021/03/03 to 2021/08/09 in Ecuador, collected by Segurilab and Previne Salub laboratories [4]. It consists of 375 observations (clinical cases) with 24 variables, including demographic variables (sex, age), CBC results for each patient (including leukocytes counts, WBC counts and percentages, mcv, mch, etc.) and their infection status. Table 1 shows all variables in our dataset with brief descriptions.

Variable names	Description
sex	Sex of patient
age	Patient's age
leukocytes	Leukocytes counts (per μL)
neutrophilsP	Neutrophils percentage (%)
lymphocytesP	Lymphocytes percentage (%)
monocytesP	Monocytes percentage (%)
eosinophilsP	Eosinophils percentage (%)
basophilsP	Basophils percentage (%)
neutrophils	Neutrophils counts (per μL)
lymphocytes	Lymphocytes counts (per μL)
NLR	Neutrophil/Lymphocyte ratio
monocytes	Monocytes counts (per μL)
eosinophils	Eosinophils counts (per μL)
basophils	Basophils counts (per μL)
redbloodcells	red blood cells counts (per μL)
mcv	Mean corpuscular volume (fL)
mch	Mean corpuscular hemoglobin (pg)
mchc	Mean corpuscular hemoglobin concentration (mg/dl)
rdwP	Red Blood Cell Distribution Width (%)
hemoglobin	Hemoglobin (mg/dl)
hematocritP	Hematocrit percentage (%)
platelets	Platelets (mm^3)
mpv	Mean platelet volume
Condition	Positive or negative results for Covid-19

Table 1: Variables with brief descriptions of the Ecuador CBC database.

3.2 Demographic variables

Table 2 shows some summary statistics of demographic variables (sex and age) before further processing:

Dataset	Samples	PCR Positive			PCR Negative		
ECU	Selected	Male	Female	Total	Male	Female	Total
Sample size	375	75	49	124	160	91	251
Mean age	36.86	38.41	35.00	37.06	37.09	36.18	36.76
Standard deviation	12.84	14.22	14.33	14.3	11.26	13.46	12.08

Table 2: Demographic data summary for ECU raw dataset.

Over all, we can see that 62.67% of the clinical cases are males while 37.33% of the clinical cases are females. Also, 33.07% of these clinical cases test positive while 66.93% of these clinical cases test negative.

3.3 Blood variables

Some boxplots of all the blood variables before further processing are shown in Figure 3 in their original scale:

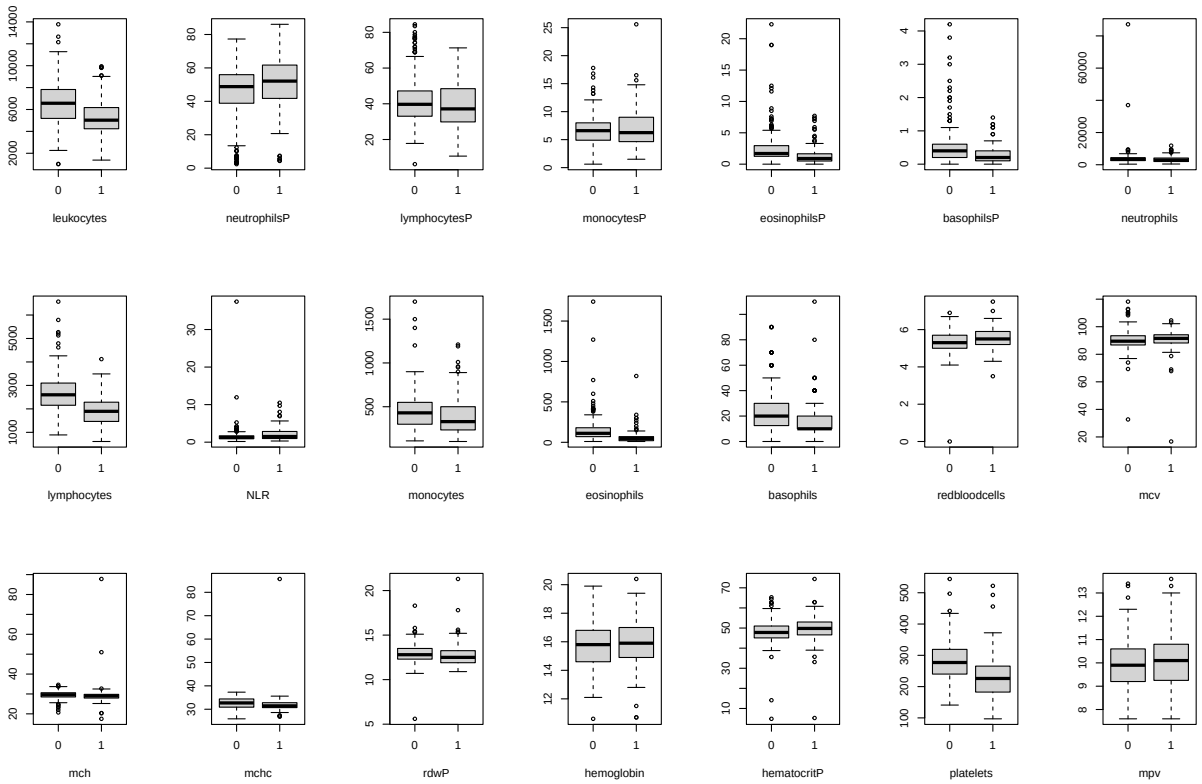

Figure 3: Boxplots of all the blood variables across different condition groups.

From Figure 3, we can see that some variables are positive-associated with COVID condition, including neutrophilsP and mpv. These variables have relatively higher levels in the positive group compared to the negative group. Also, we can see that some variables are negative-associated with COVID condition, including leukocytes, lymphocytes, eosinophils, basophils, platelets, etc. These variables have relatively higher levels in the negative group compared to the positive group.

3.4 Data cleaning

One thing to notice is that our dataset contains redundant information. For example, we are given the cell counts data of neutrophils, lymphocytes, monocytes, eosinophils, basophils and their summation data along with percentage data. Also, we are given the neutrophil/lymphocytes ratio data (NLR) despite the fact that we can calculate these values directly. Based on this observation, some of our variables should be excluded before we perform preliminary PCA on all possible predictors and try to avoid collinearity problems. Another thing to notice is that our sex variable is a categorical variable that we should exclude in our PCA as well.

Another problem is that we find the leukocytes counts data do not match the summation of five WBC counts data. And this problem cannot be fully explained by the presence of other

white blood cells such as plasma cells, as we observe the summation of five WBC counts data exceeds leukocytes counts data in some cases (see the top of Figure 4 and Figure 5). To deal with this problem, we established contact with Prof. Márcio Dorn, one of the original data generators and maintainer of the Ecuador CBC database [4]. According to his professional perspective, the presence of a large deviance of between the summation and the leukocytes counts indicates that special medical conditions may have happened. For example, certain viral infections can cause a decrease in circulating lymphocytes, while an intense allergic reaction can result in a significant increase in eosinophils. These changes can cause discrepancies in the total numbers. In the case of this data, all the patients were symptomatic and we want to exclude those cases for an accurate compositional data transformation.

Based on the observation of redundant variables and the deviance between leukocytes counts and summation of WBC, we first apply some data cleaning procedure to our dataset. For patients whose leukocytes counts and WBC sum has a relative deviance larger than 10%, we exclude those cases and get an updated Table 3 of summary statistics and deviance figures (see the bottom of Figure 4 and Figure 5).

Dataset	Samples	PCR Positive			PCR Negative		
ECU	Selected	Male	Female	Total	Male	Female	Total
Sample number	329	65	42	107	146	76	222
Mean age	37.16	39.49	36.19	38.2	36.92	36.16	36.66
Standard deviation	12.79	13.96	14.46	14.18	11.29	13.49	12.07

Table 3: Data Summary for ECU Cleaned Dataset.

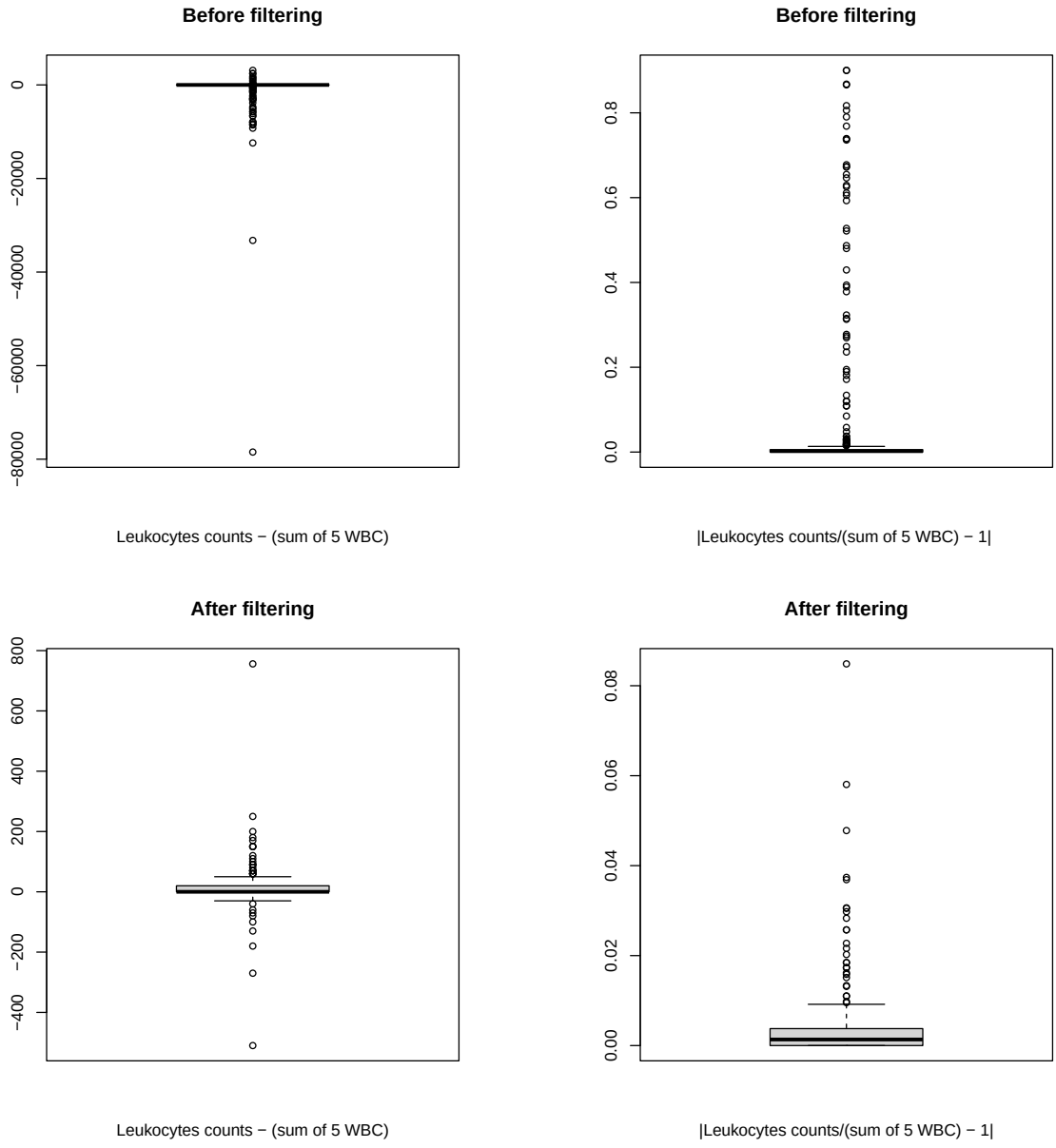

Figure 4: Boxplots of absolute and relative deviance before and after filtering.

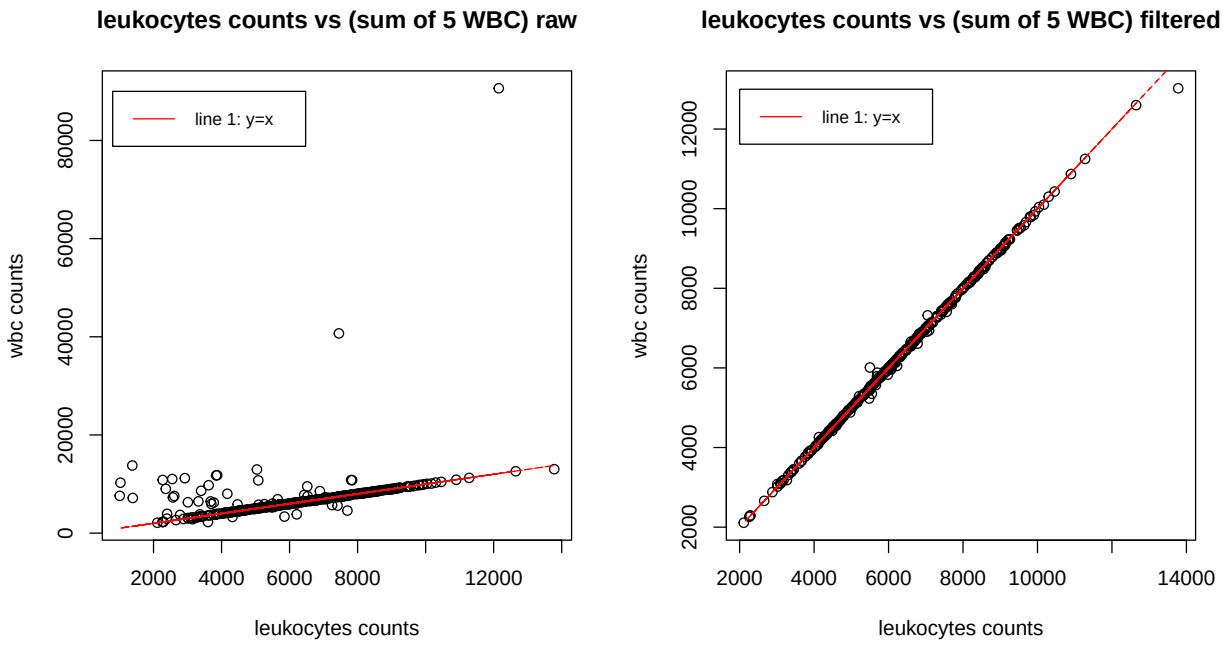

Figure 5: Scatterplots of WBC summation vs leukocytes counts before and after filtering.

It calls to the attention that 67.48% of the clinical cases is PCR-negative in our cleaned dataset, whereas all of them were symptomatic before taking the blood test and PCR test. This issue of low-sensitivity of the PCR test will be discussed in greater details in the discussions and conclusions section.

4 Data Analysis

In this section, we apply PCA, LDA, LR-PCA, and LR-LDA methods to the Ecuador CBC dataset and assess their performances.

4.1 PCA

4.1.1 PCA using all variables

Before the principal component analysis, we exclude the following variables because of redundant information issue mentioned earlier: leukocytes, NLR, and percentage variables of five major WBC. Moreover, sex variable is not used for the PCA as it is binary. The condition variable (RT-qPCR) is also excluded as it is our target outcome. Employing the PCA technique to all the other variables of the Ecuador CBC data gives us a biplot shown below (see Figure 6). This biplot excludes one extreme point/outlier, in order to get a reasonable visualization of all the other observations.

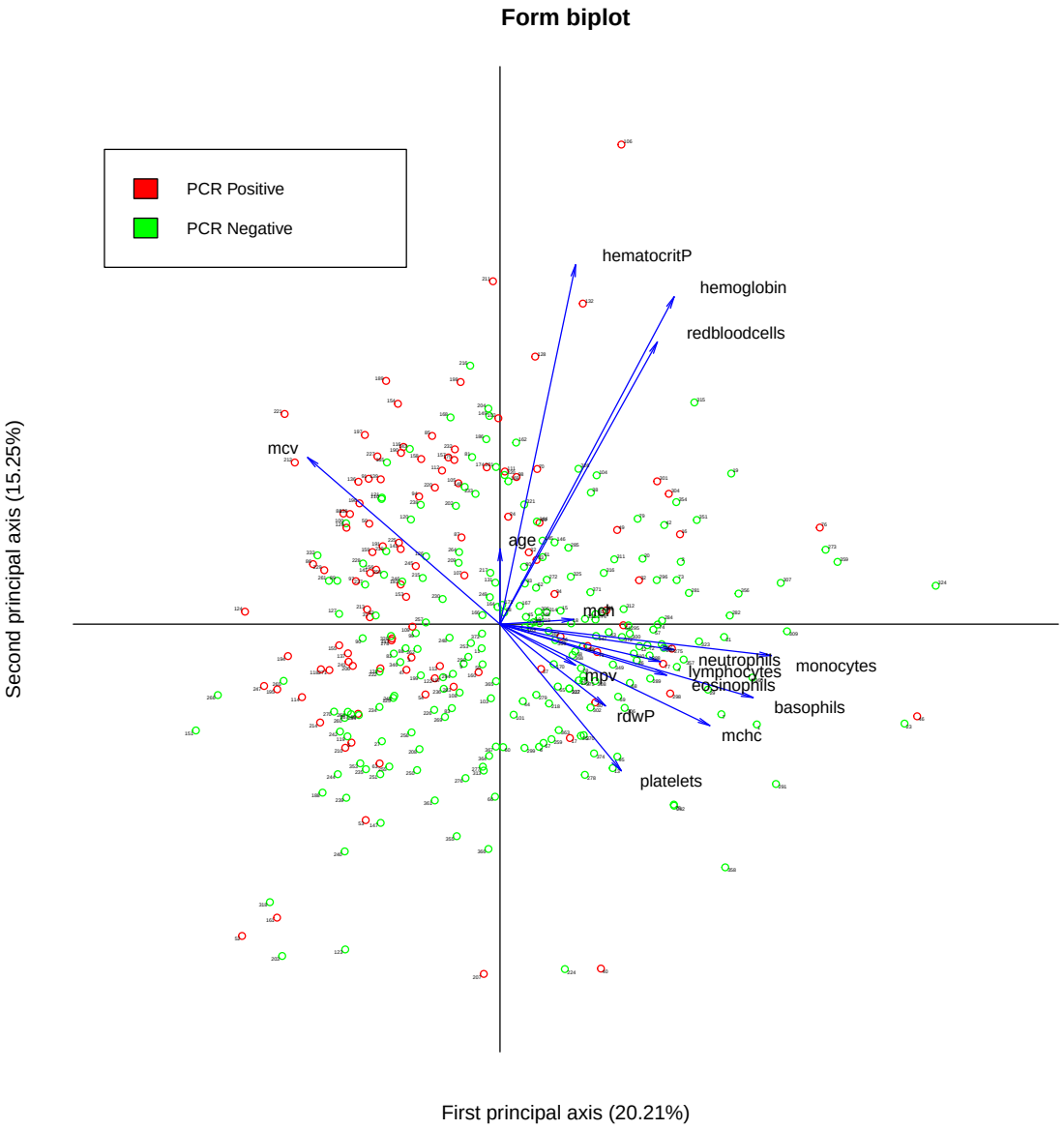

Figure 6: Preliminary PCA using all variables.

In the above biplot, the cases with COVID-19 positive condition are colored as red points while the cases with COVID-19 negative condition are colored as green points. We use the correlation matrix in our PCA since we have many variables with incomparable units. Also, we use the form biplot for final demonstration. According to this data-visualization, our first principal component seems to be an important feature for COVID-19 condition classification, as most of the positive cases located on the left hand side of our figure while negative cases located on the right. And this component is negatively associated with the mcv variable and positively associated with all the other variables. However, the first two principal components only explain 35.46% of total variance, suggesting only using two principal components may be insufficient to capture most variance of our dataset. We need eight principal components for our selected features if we aim to explain more than 80% of total variance (see scree plot Figure 7).

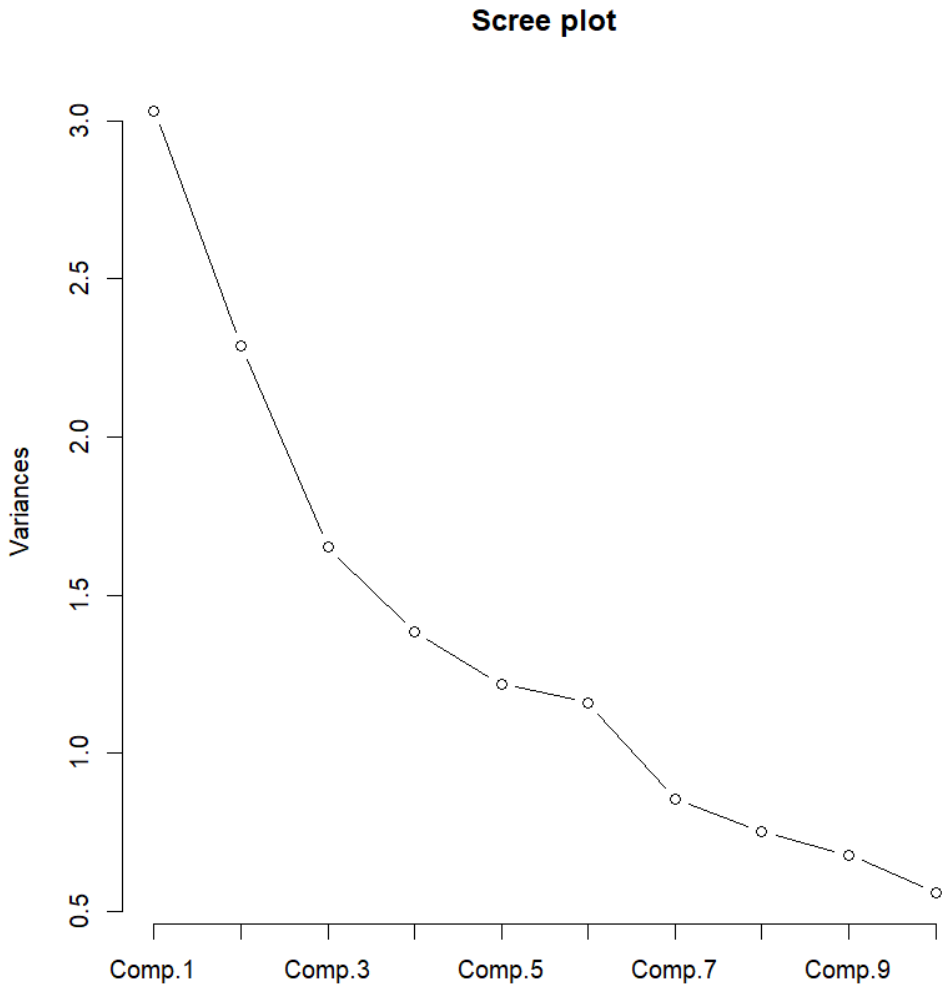

Figure 7: Scree plot of PCA using all variables.

4.1.2 PCA using WBC counts variables

Considering the prior knowledge of the diagnostic power of the five major WBC counts, we reemploy the PCA technique to the WBC counts dataset to give us a second form biplot (see Figure 8). In our PCA, we only include neutrophils, lymphocytes, monocytes, eosinophils and basophils cell counts as our predictor variables.

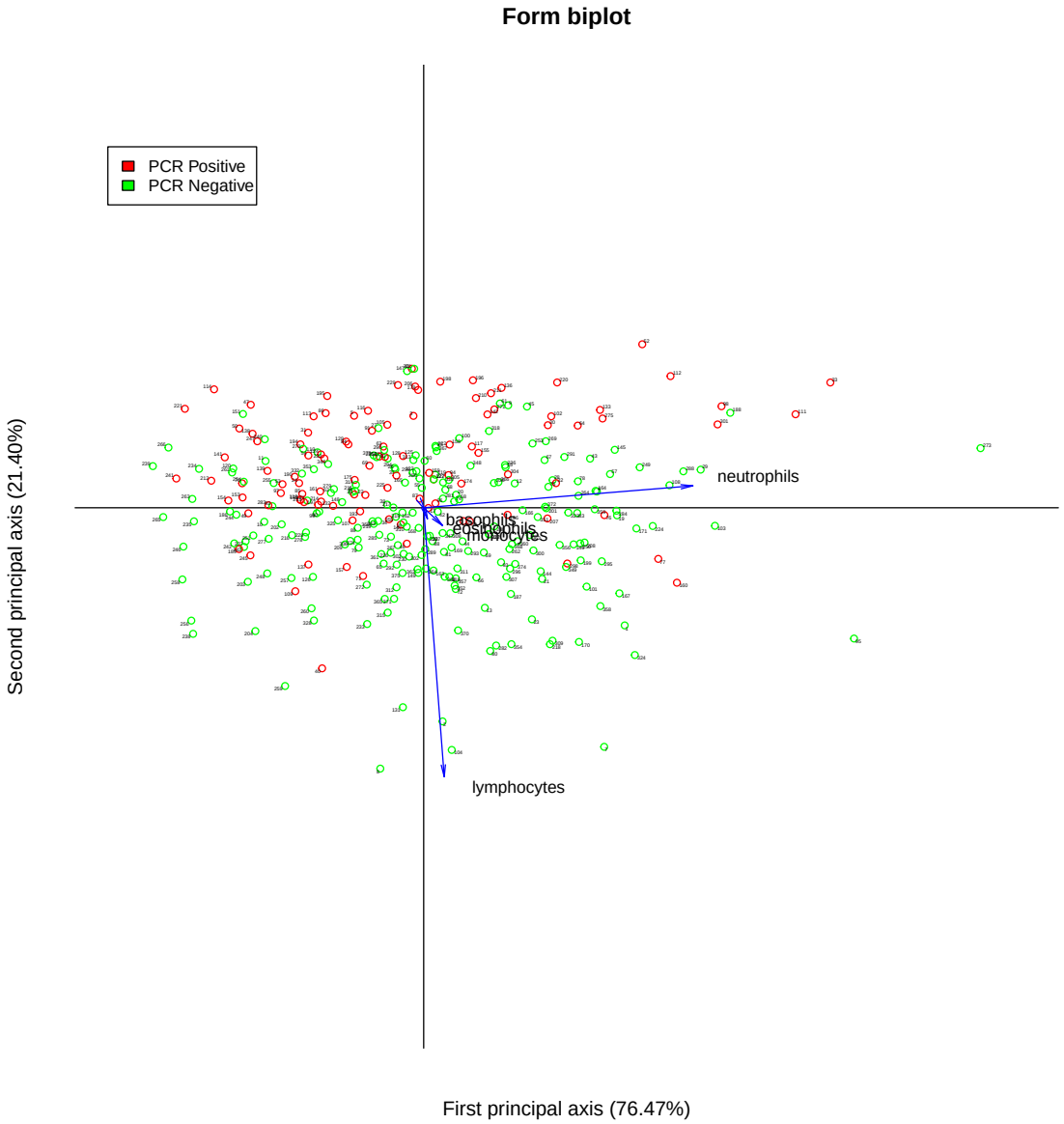

Figure 8: Form biplot of PCA using WBC counts data only.

In this setting, we are using the covariance matrix in our PCA first since we are dealing with count variables of all predictors and their units are perfectly comparable with each other. According to this data-visualization, our second principal component seems to be an important feature for COVID-19 condition classification, as most of the positive cases are located at the top of our figure

while negative cases located at the bottom. And this component is highly correlated with the lymphocytes variable and almost uncorrelated with all the other variables. Moreover, the first two principal components already explain 97.87% of total variance, suggesting using two principal components is sufficient to capture most of our white cell count dataset's structure (see scree plot Figure 9). One noteworthy observation is that there still exists mixing of positive and negative cases (e.g. cases 40,77,160, and 188) since the PCA method is not optimized for class separation. This mixing may be attributed to factors such as the presence of other respiratory diseases or variations in the stages of the immunological window. Further details are discussed in the Discussion and Conclusions section.

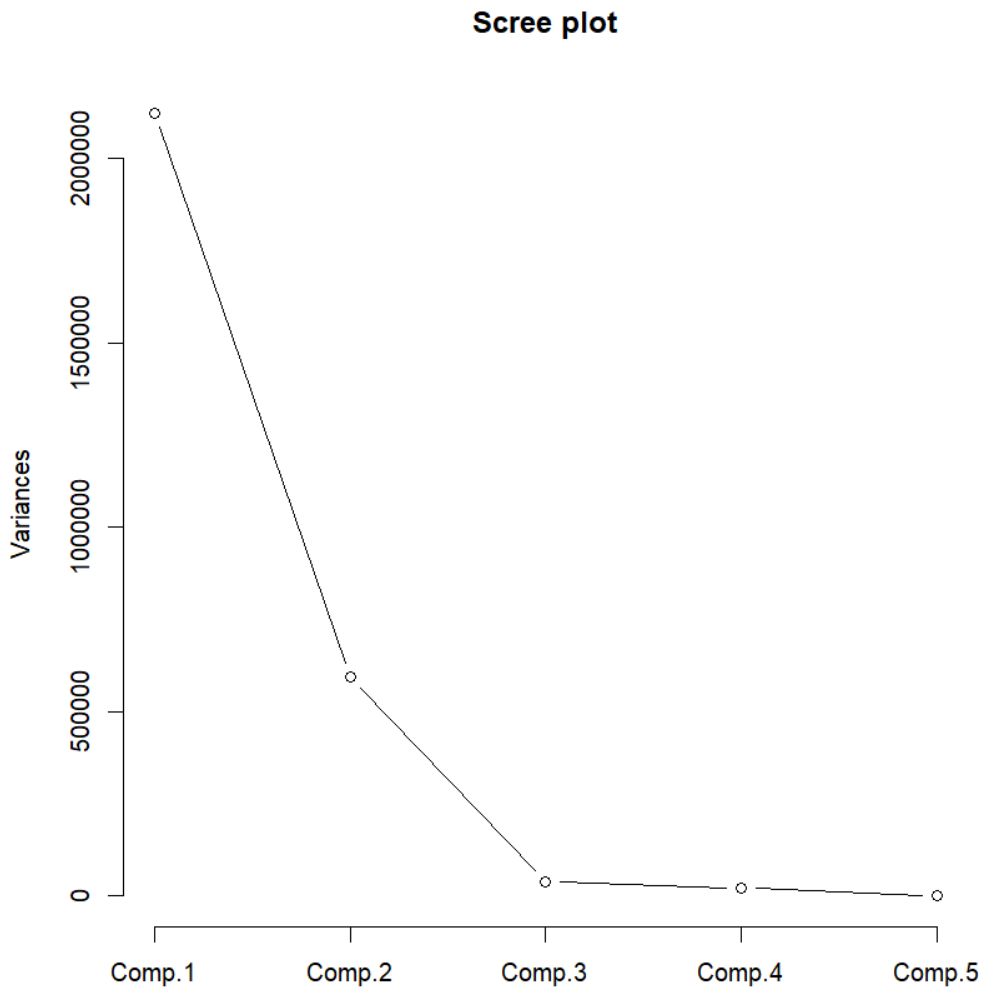

Figure 9: Scree plot of PCA using WBC counts data only.

Moreover, we can plot a covariance biplot (also based on covariance matrix instead of correlation matrix) to visualize the correlation structure of components (see Figure 10).

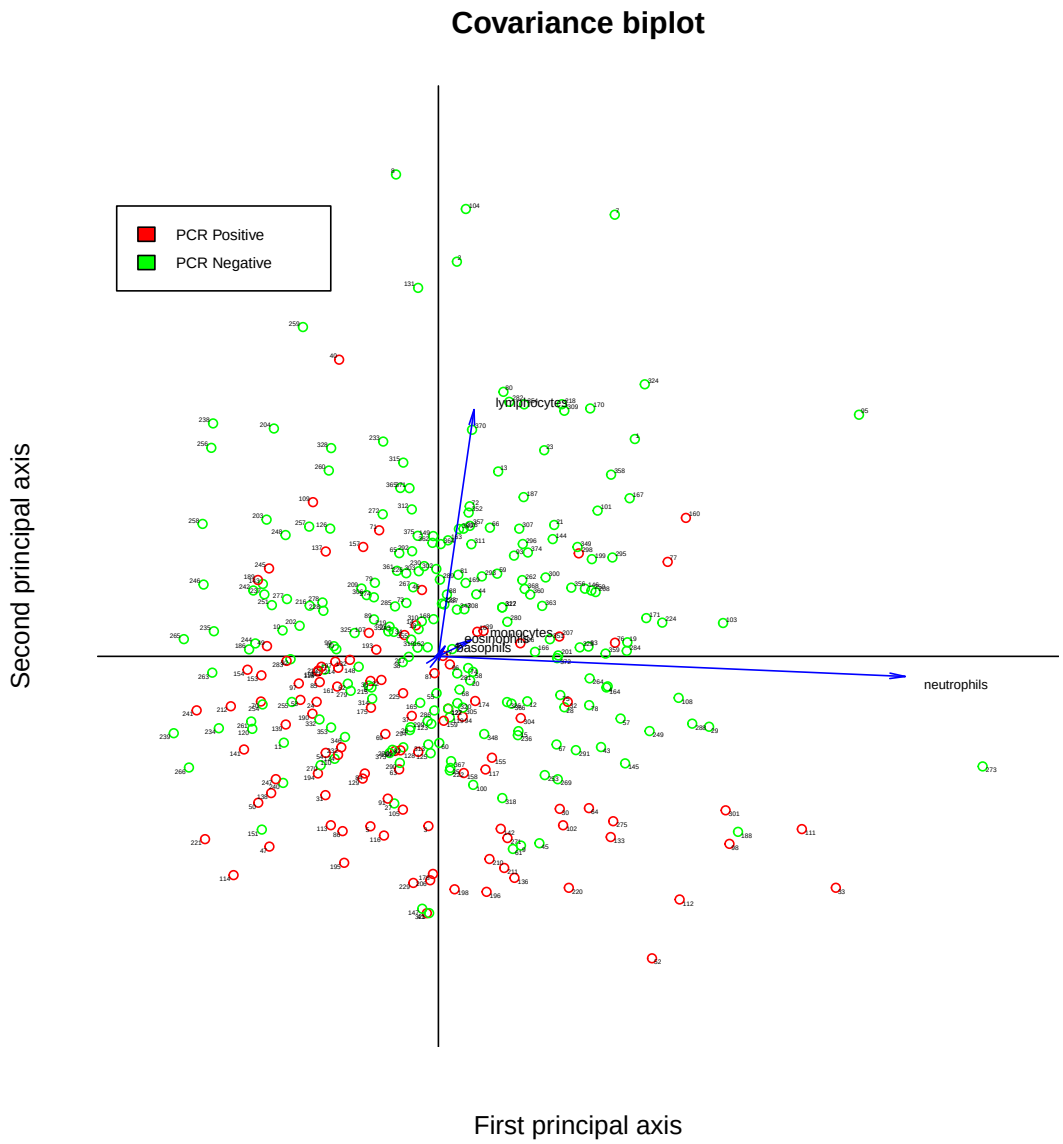

Figure 10: Covariance biplot of PCA using WBC counts data only.

From both the form biplot and covariance biplot, we can see that lymphocytes and neutrophils dominate the first two PCs. They have the largest counts and also the largest variances. To reduce the influence of the large count number of lymphocytes and neutrophils, we could also standardize the data and make a correlation based form biplot as shown in Figure 11. In this correlation based form biplot, the first two principal components only explain 63.80% of total variance (see Figure 12).

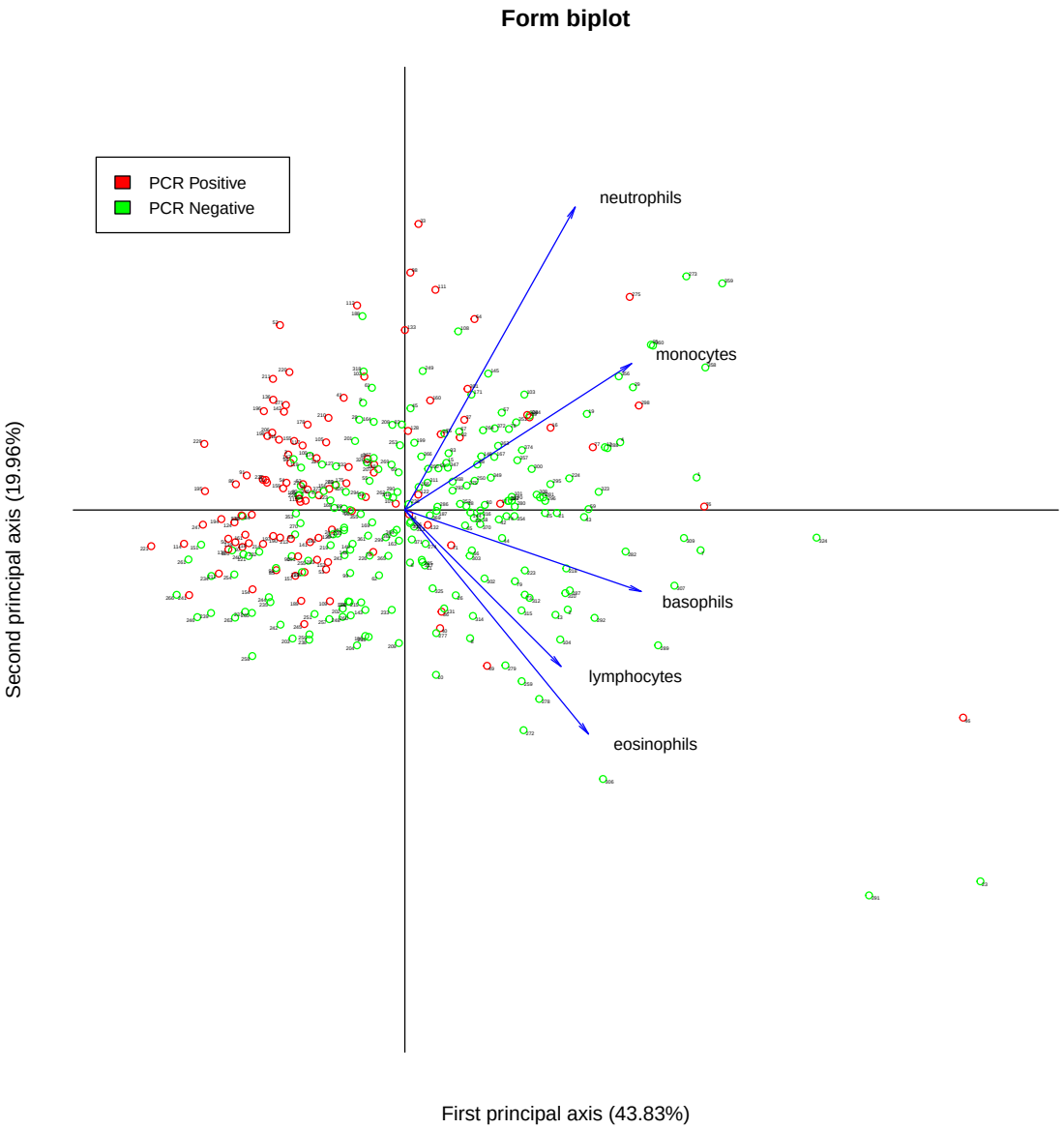

Figure 11: Correlation-based form biplot of PCA using WBC counts data.

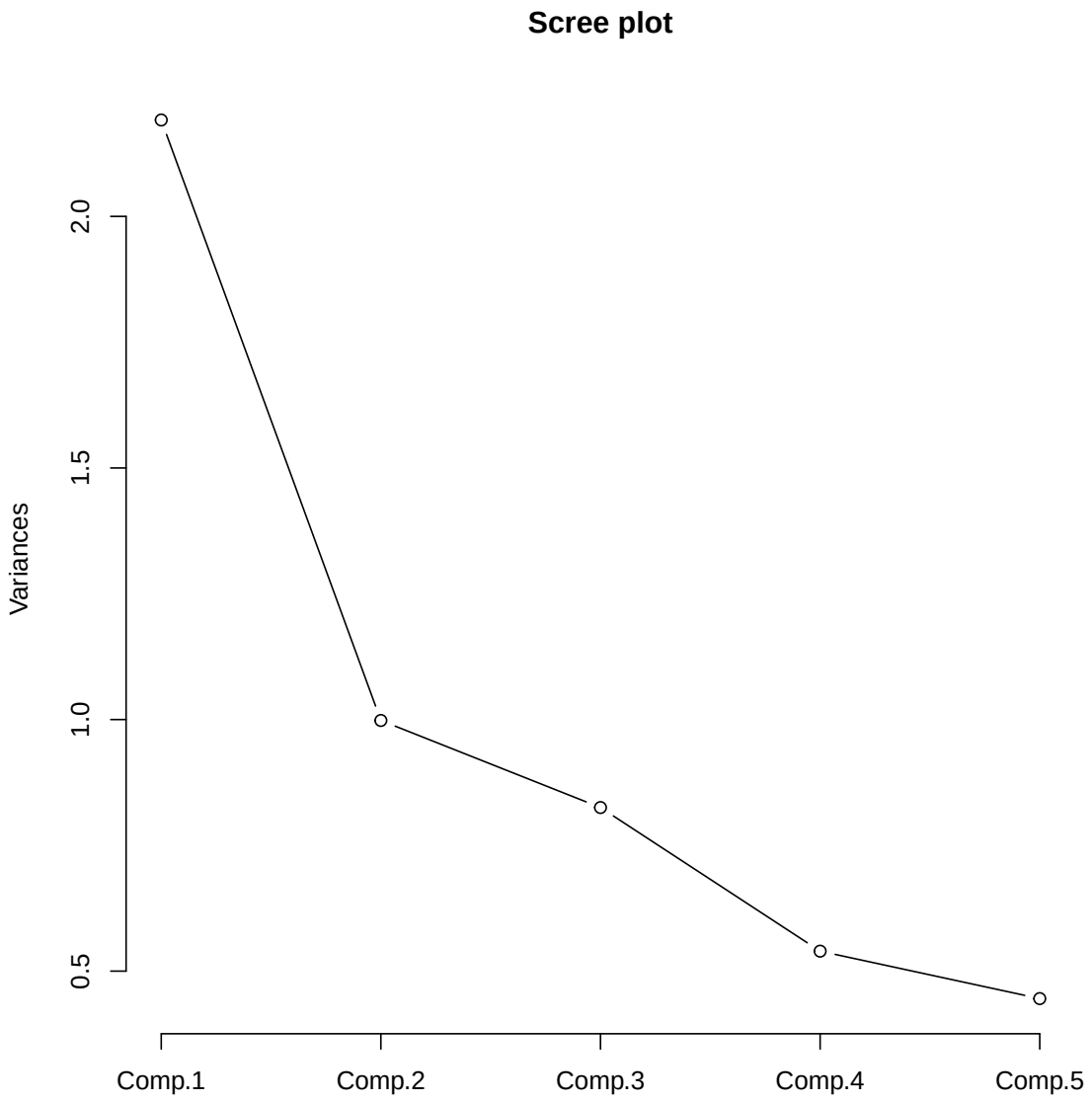

Figure 12: Scree plot of correlation based PCA using WBC counts.

From Figure 11, we can see that now five counts variables have similar length and all five of them are positive correlated with the first principal component. Since most of PCR positive cases are located on the left half of the biplot and most PCR negative cases are located on the right half of the biplot, it seems that all five WBC counts variables are positively correlated with the PCR negative status in our standardized WBC counts dataset. Thus, in this analysis, we can interpret the first principal component as the total WBC count and the second principal component as the neutrophils-eosinophils difference, with the total count have a high predictive power for COVID-19 status classification. This information is not evident from the covariance-based biplots, as lymphocytes and neutrophils dominate the first two PCs.

4.2 LDA

4.2.1 LDA using all variables

We apply LDA to the full dataset and WBC counts dataset before the log-ratio transformation. Figure 13 shows LDA with the *standardized* full dataset. Notice that standardization won't give us different classification results because of the invariance property of LDA, and we standardize our full dataset in order to get comparable linear discriminant coefficients for all variables.

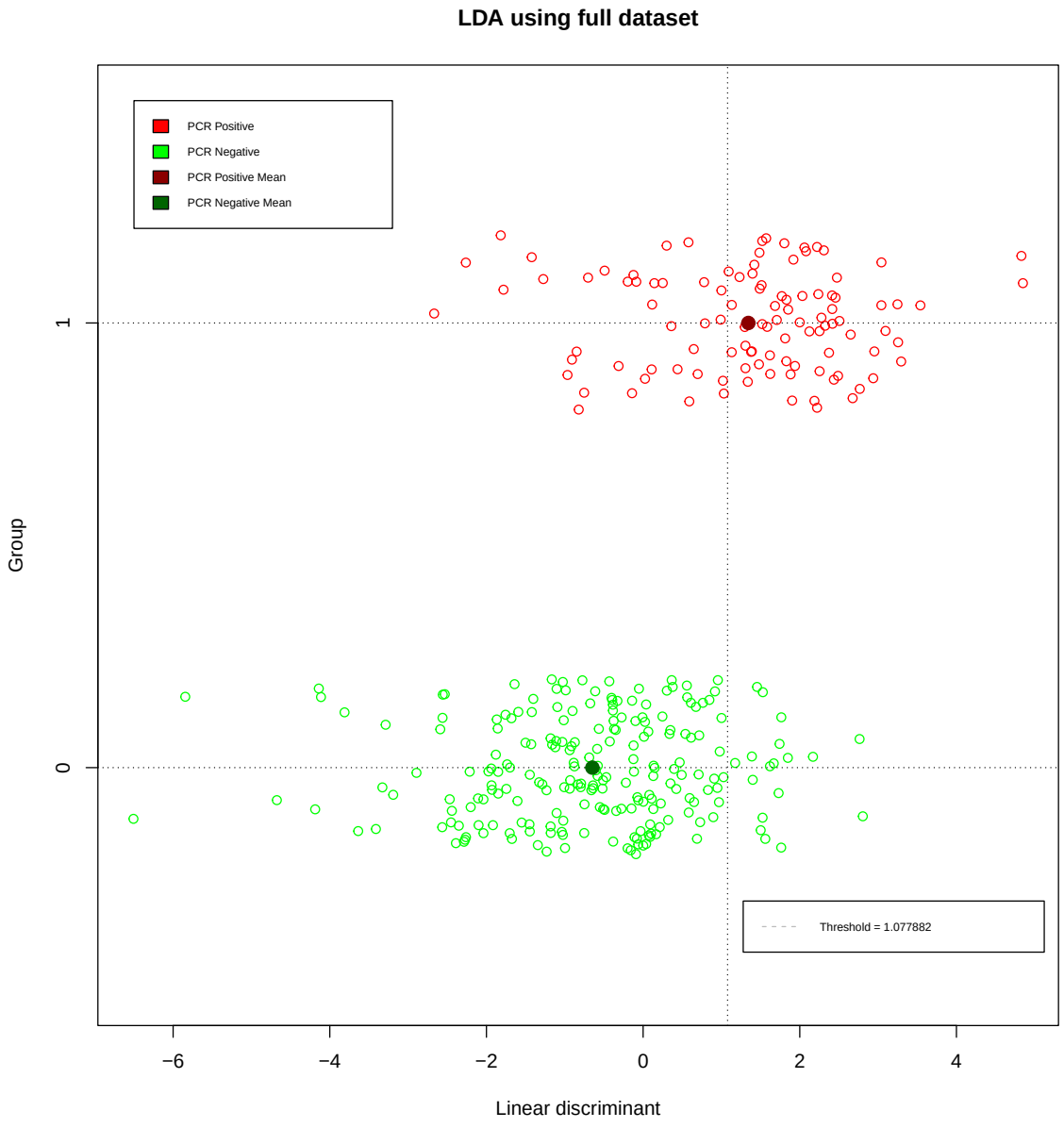

Figure 13: A demonstration of LDA using all variables.

One thing to notice is that we add a small random noise (through the "jitter" function in R) to the COVID condition variable prior plotting the results to reduce the amount of overlapping

points. In this Figure 13, we are classifying the outcome without cross-validation to obtain a classifier threshold $y_{\text{classifier}} := 1/2(\bar{y}_0 + \bar{y}_1) + ln(p_0/p_1) = 1.0779$ (plotted as the vertical dotted line, with equal cost of false positive and false negative). Data points on the left hand side of the classifier are classified as PCR negative and data points on the right hand side are classified as PCR positive. From Figure 13 we can see that the LDA technique performs better for classifying PCR negative patients compared to PCR positive patients, as more red points are misclassified to the left hand side of the classifier compared to green points on the right hand side of the classifier. This fact can also be seen by the confusion matrix of LDA on the full dataset (see Table 4, with $18 < 36$). We obtain a classification rate of 83.587% using LDA without cross-validation.

	Prediction	
True	0	1
0	204	18
1	36	71

Table 4: Confusion matrix for LDA with the full dataset.

We present the coefficient vector $\mathbf{a}$ of linear discriminant function $y_i = \mathbf{a}^T \cdot \mathbf{x}_i$ in Table 5 and Figure 14. In Table 5, the variables with a positive coefficient sign indicate that those variables are positively associated COVID-19 condition, while those variables with a negative coefficient sign are negatively associated with COVID-19 condition. From both the table and the bar plot, we can see that the variable lymphocytes has the most prediction power. Red blood cells, platelets, and eosinophils also have relatively large prediction power, as they have coefficients with a large absolute value. One thing to notice is that, according to the linear discriminant function coefficient, mcv doesn't have much more prediction power compared to other variables and this observation is different from the the PCA biplot using all variables shown in Figure 6. The reason behind this inconsistency could be due to the different targets of PCA and LDA techniques. Another surprising thing to notice is that monocytes count variable has a different direction compared to all the other four WBC count variables (neutrophils, lymphocytes, eosinophils and basophils) in the LDA. However, in the principal component analysis all five WBC counts variables have similar directions (see Figure 6). This inconsistency could be due to several reasons that need further investigation. For example, the reason could be statistic. The compositional data structure require WBC counts variables to sum up to close total counts, and this collinearity problem could result in the constraint of their LDA coefficient making the direction of monocytes variable to flip. Also, a hidden biological mechanism linking monocytes counts variable with COVID-19 condition may cause the sign of monocytes to flip when it comes to COVID-19 prediction instead of just data visualization.

age	-0.0238
neutrophils	-0.0358
lymphocytes	**-1.0860**
monocytes	0.1695
eosinophils	**-0.5217**
basophils	-0.1480
redbloodcells	**0.5700**
mcv	-0.2060
mch	0.0910
mchc	-0.2783
rdwP	0.0766
hemoglobin	-0.2579
hematocritP	0.2619
platelets	**-0.5665**
mpv	0.1437

Table 5: Coefficient vector of the linear discriminant function (normalized full dataset).

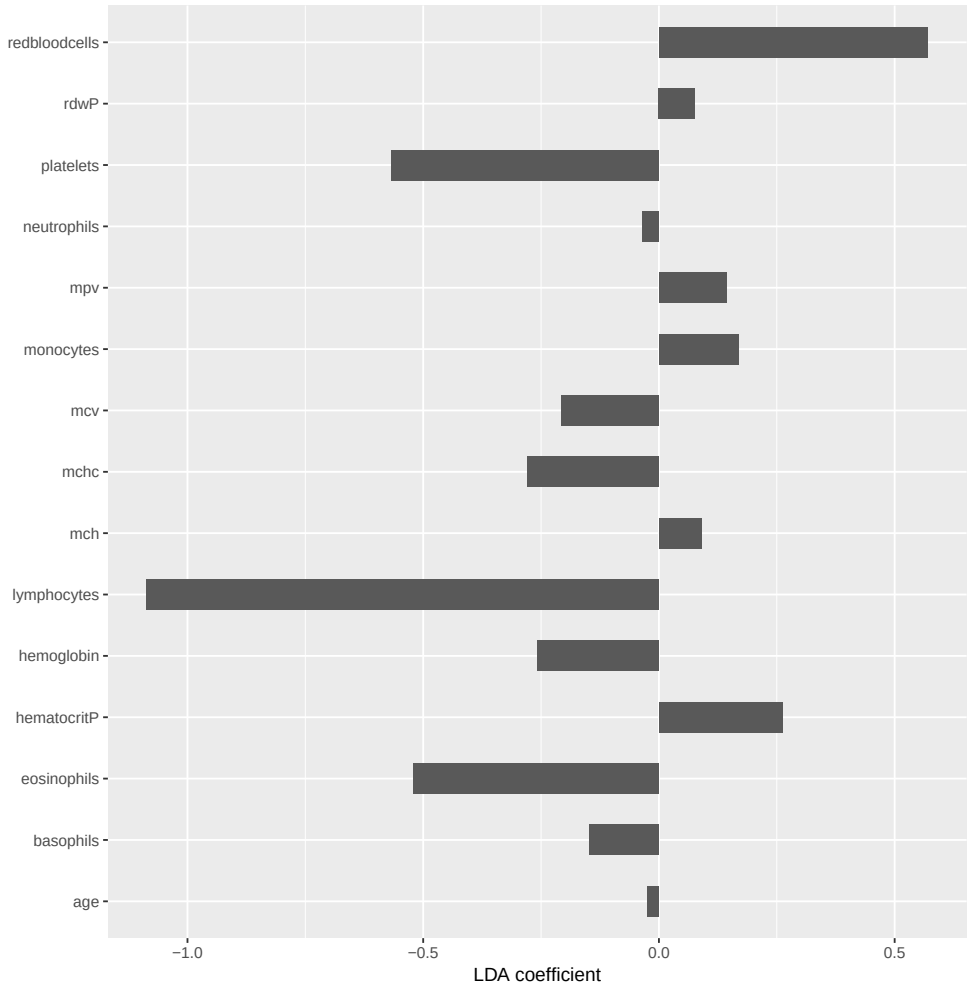

Figure 14: LDA coefficients using all variables.

The receiver operating characteristic curve (ROC curve) for LDA using all variables is shown in the combined ROC Figure 17. We use the ROC function from the R package pROC to generate all the ROC figures in this thesis[20]. The area under the curve (AUC) is 0.8487 without cross-validation.

The classification rate is calculated using leave one out (LOO) cross validation. We obtain a classification rate of 81.459% with the corresponding confusion matrix (see Table 6)

	Prediction	
True	0	1
0	200	22
1	39	68

Table 6: Cross-validated confusion matrix for LDA with full dataset.

4.2.2 LDA using WBC counts variables

We repeat the LDA analysis using only the WBC counts dataset. Figure 15 demonstrates LDA with the WBC dataset. Since all five WBC counts variables have the same unit, there is no need

to standardize our dataset this time.

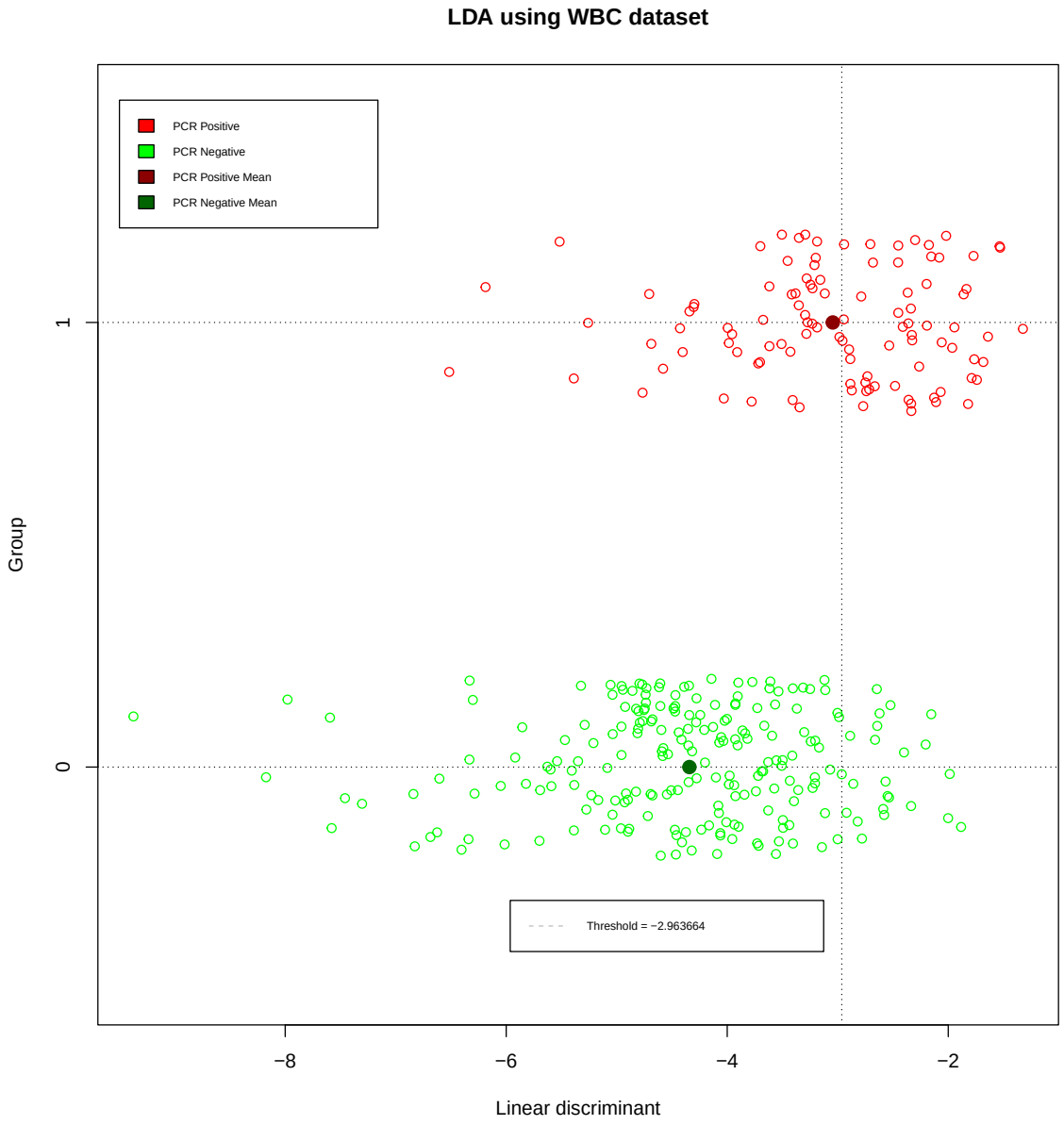

Figure 15: LDA using five major WBC counts variables.

We still add a small random noise (through the "jitter" function in R) to the COVID condition variable prior plotting the results to reduce the amount of overlapping points. In Figure 15, we are classifying outcome without cross-validation to obtain a classifier threshold $y_{\text{classifier}} := 1/2(\bar{y}_0 + \bar{y}_1) + ln(p_0/p_1) = -2.9637$ (plotted as the vertical dotted line, with equal cost of false positive and false negative). Data points on the left hand side of the classifier are classified as PCR negative and data points on the right hand side are classified as PCR positive. From Figure 15 we can see that LDA technique still performs better for classifying PCR negative patients compared to PCR positive patients in WBC dataset, as more red points are misclassified to the left hand side of classifier compared to green points on the right hand side of classifier. Moreover, the LDA technique

performs worse in the WBC dataset compared to the full dataset as we have more misclassified cases. Those facts can also be seen by the confusion matrix of LDA on the WBC dataset (see Table 7). We obtain a classification rate of 77.508% using LDA without cross-validation.

	Prediction	
True	0	1
0	199	23
1	51	56

Table 7: Confusion matrix for LDA with WBC dataset.

We present the coefficient vector $\mathbf{a}$ of linear discriminant function $y_i = \mathbf{a}^T \cdot \mathbf{x}_i$ for WBC dataset in Table 8 and Figure 16. Again, we have variables with a positive coefficient sign indicating that those variables are positively associated with the COVID-19 condition, while variables with a negative coefficient sign indicate that they are negatively associated with COVID-19 condition. From both the table and the bar plot, we can see that the basophils variable has the most prediction power. This observation is hard to be seen from the PCA biplot in Figure 8 as well, and could be due to the different targets of PCA and LDA techniques.

neutrophils	-0.0001
lymphocytes	-0.0014
monocytes	0.0010
eosinophils	-0.0025
basophils	**-0.0137**

Table 8: Coefficient vector of the linear discriminant function (WBC dataset).

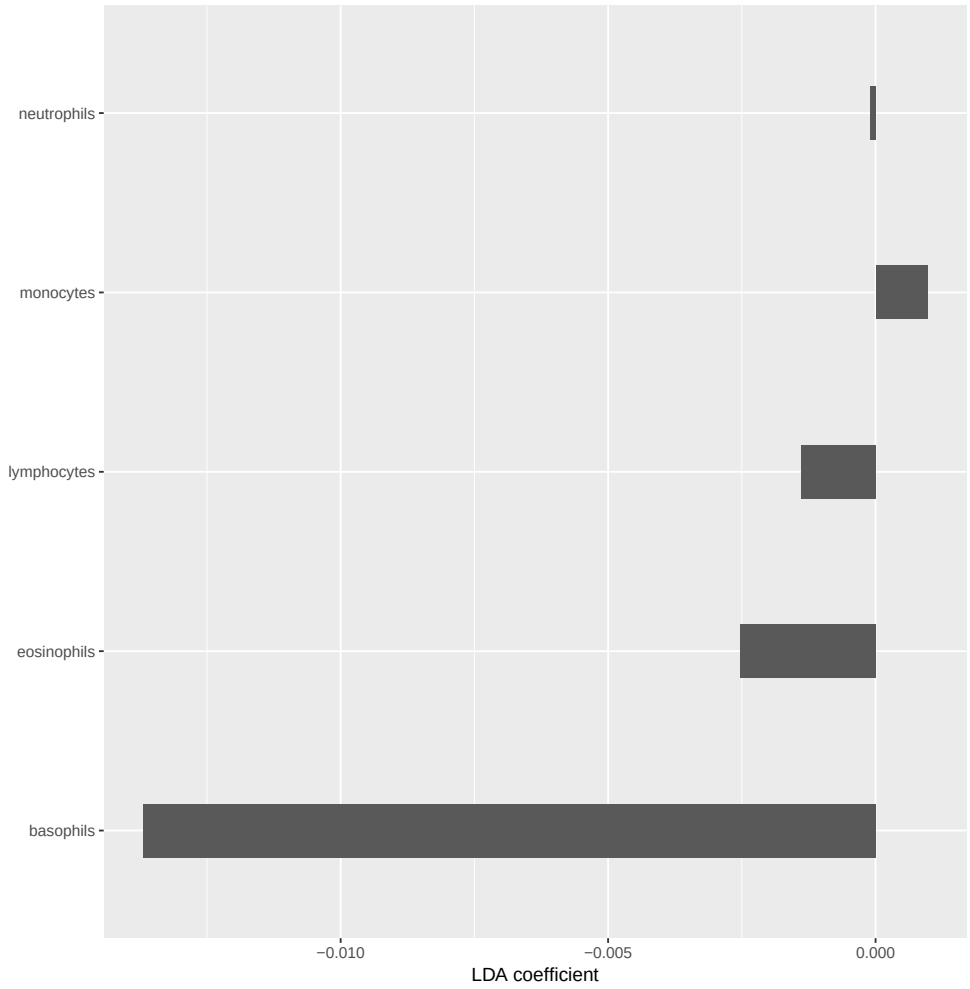

Figure 16: A bar plot for LDA coefficients using WBC counts variables.

The receiver operating characteristic curve (ROC curve) for LDA using WBC counts variables is shown in the combined ROC Figure 17. The area under the curve (AUC) is 0.8129 without cross-validation.

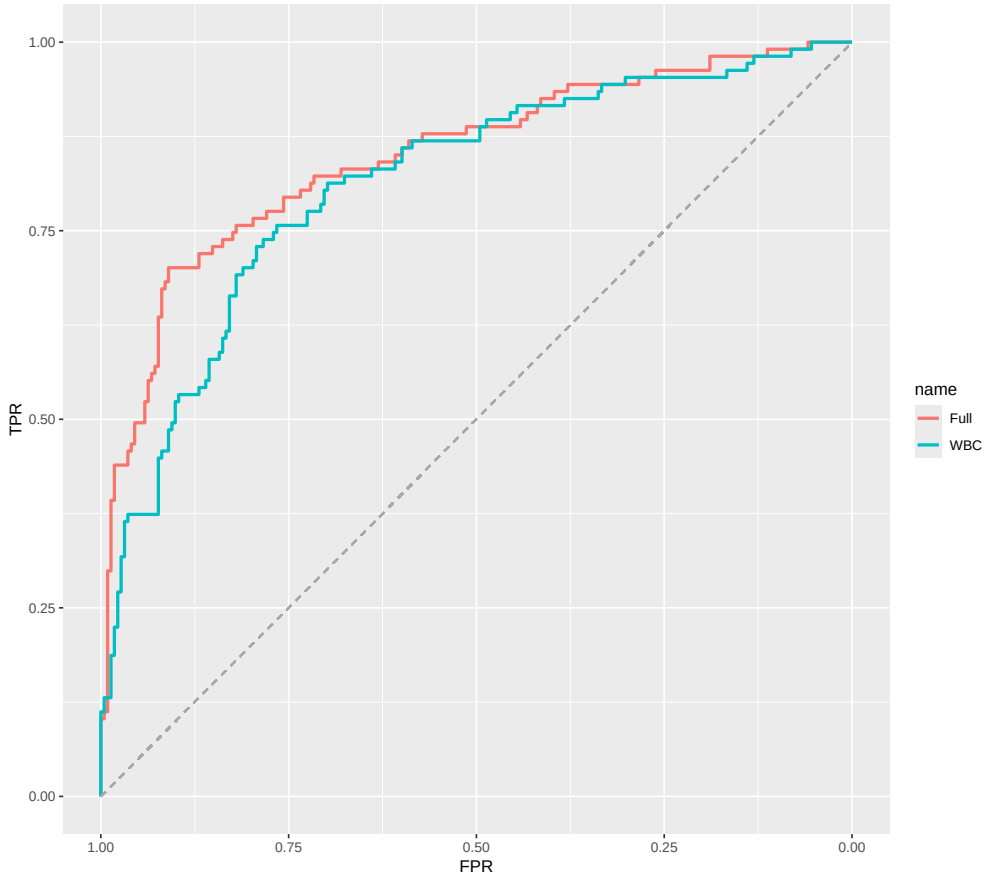

Figure 17: ROC curves for LDA using Full dataset and WBC counts dataset.

The classification rate is also calculated using LOO cross validation. We obtain a classification rate of 76.900% with the corresponding confusion matrix (see Table 9)

	Prediction	
True	0	1
0	198	24
1	52	55

Table 9: Cross-validated confusion matrix for LDA with WBC dataset.

Table 10 summarizes the performances of LDA classification models using all variables and WBC counts variables with or without cross-validation. The classification rates are more accurate with cross-validated models as they compensate for overfitting problems.

Models	Classification rate	AUC
LDA full without CV	83.587%	0.8487
LDA full with CV	81.459%	0.8159
LDA WBC without CV	77.508%	0.8129
LDA WBC with CV	76.900%	0.7989

Table 10: Summary table assessing performances of LDA models.

4.3 LR-PCA

Now we apply LR-PCA method to the WBC counts data. In order to do so, the WBC counts data is first transformed to a compositional data by dividing each entry with its row summation (the sum of that cases' WBC counts). Then we are able to apply the *clr* transformation and standard PCA to obtain a form biplot (see Figure 18). Notice that we have eight cases with zero basophils counts in our WBC counts dataset, and these zeros are replaced with a small non-zero number based on a Bayesian-multiplicative replacement by the CMULTREPL function in the ZCOMPOSITIONS R package [14].

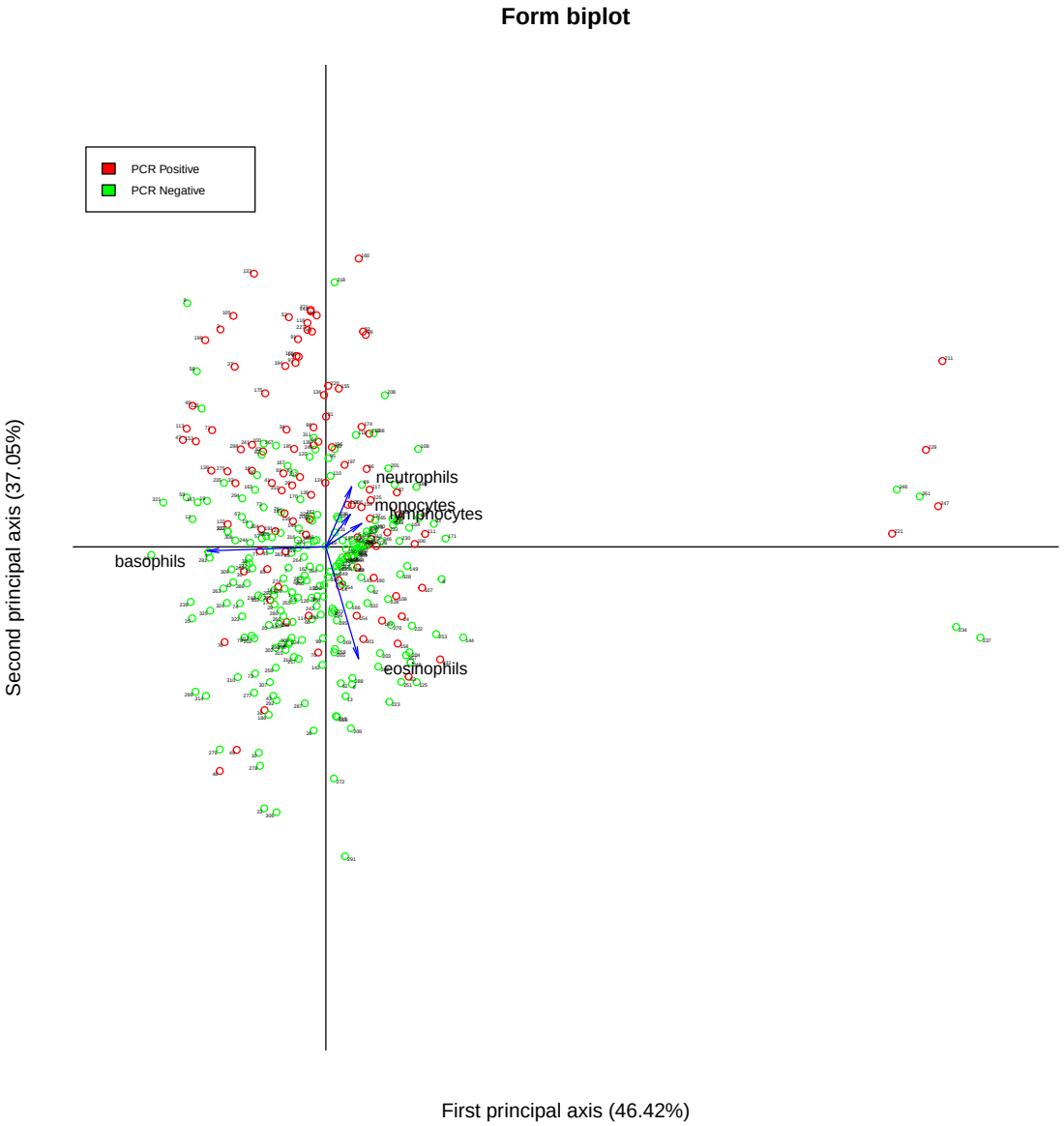

Figure 18: Form biplot of LR-PCA using WBC composition data.

According to this data-visualization, the log ratio of basophils and lymphocytes almost coincides

with the first principal axis, and the log ratio of eosinophils and neutrophils almost coincide with the second principal axis. Moreover, the log ratio of eosinophils and neutrophils is almost uncorrelated with the log ratio of basophils and lymphocytes, as the links between those vectors has an angle close to 90 degree. In the form biplot, the first two principal components explain 83.47% of the total variance, suggesting using two principal components is sufficient to capture most of *clr*-transformed WBC composition dataset's structure (see scree plot Figure 19). One thing to notice is that the minimum eigenvalue of the sample covariance matrix (see Equation (5) and (7)) is systematically zero, and this is caused by the unit-sum constraint of the compositional data. Also, we have eight extreme points in the biplot, and these coincide with eight cases having zero basophils counts.

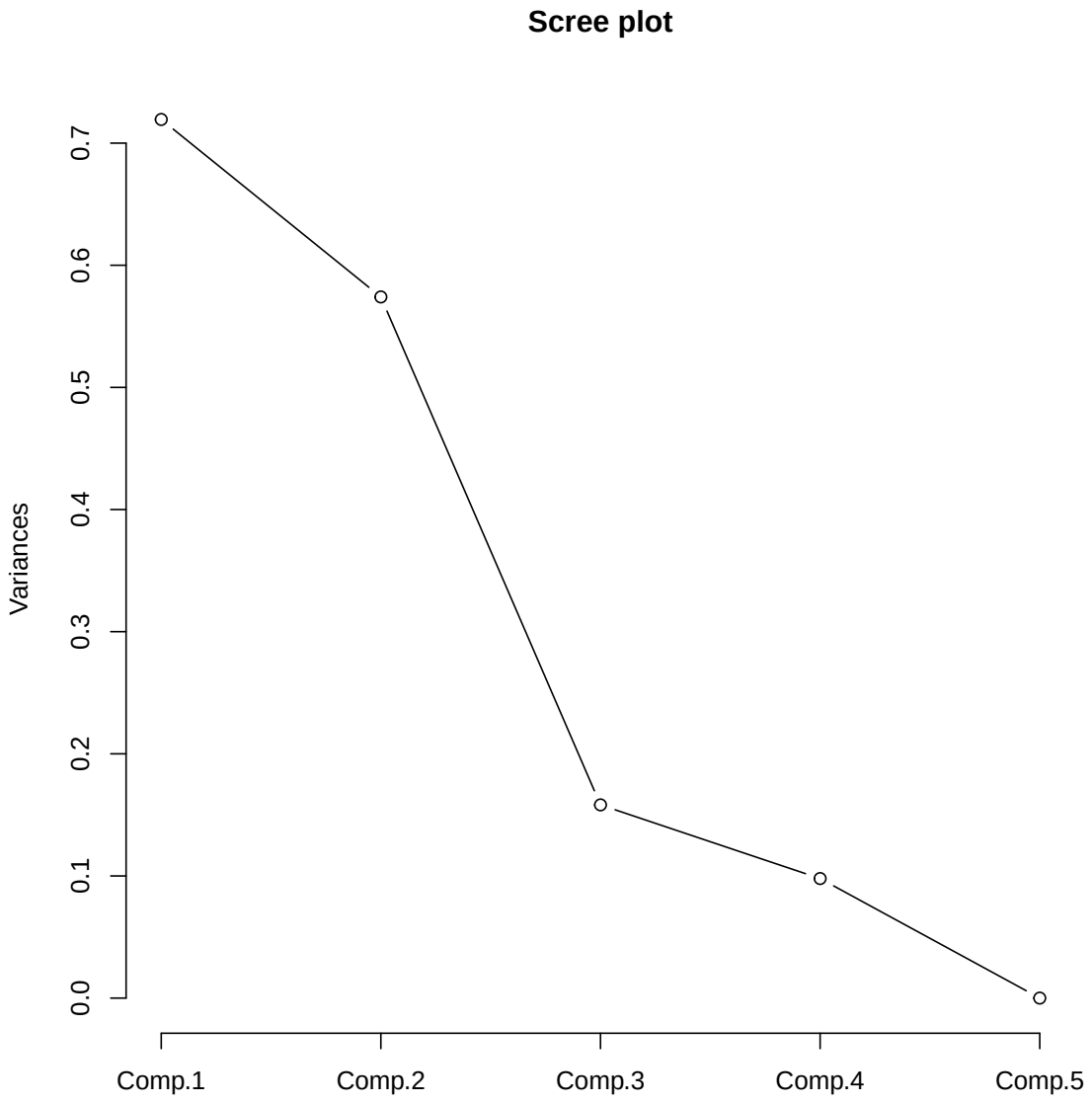

Figure 19: Scree plot of LR-PCA using WBC composition data.

Moreover, we can plot a covariance biplot (also based on covariance matrix) to visualize the correlation structure of *clr*-transformed WBC composition dataset (see Figure 20). Still, we have eight extreme points in the biplot, corresponding to eight cases with zero basophils counts.

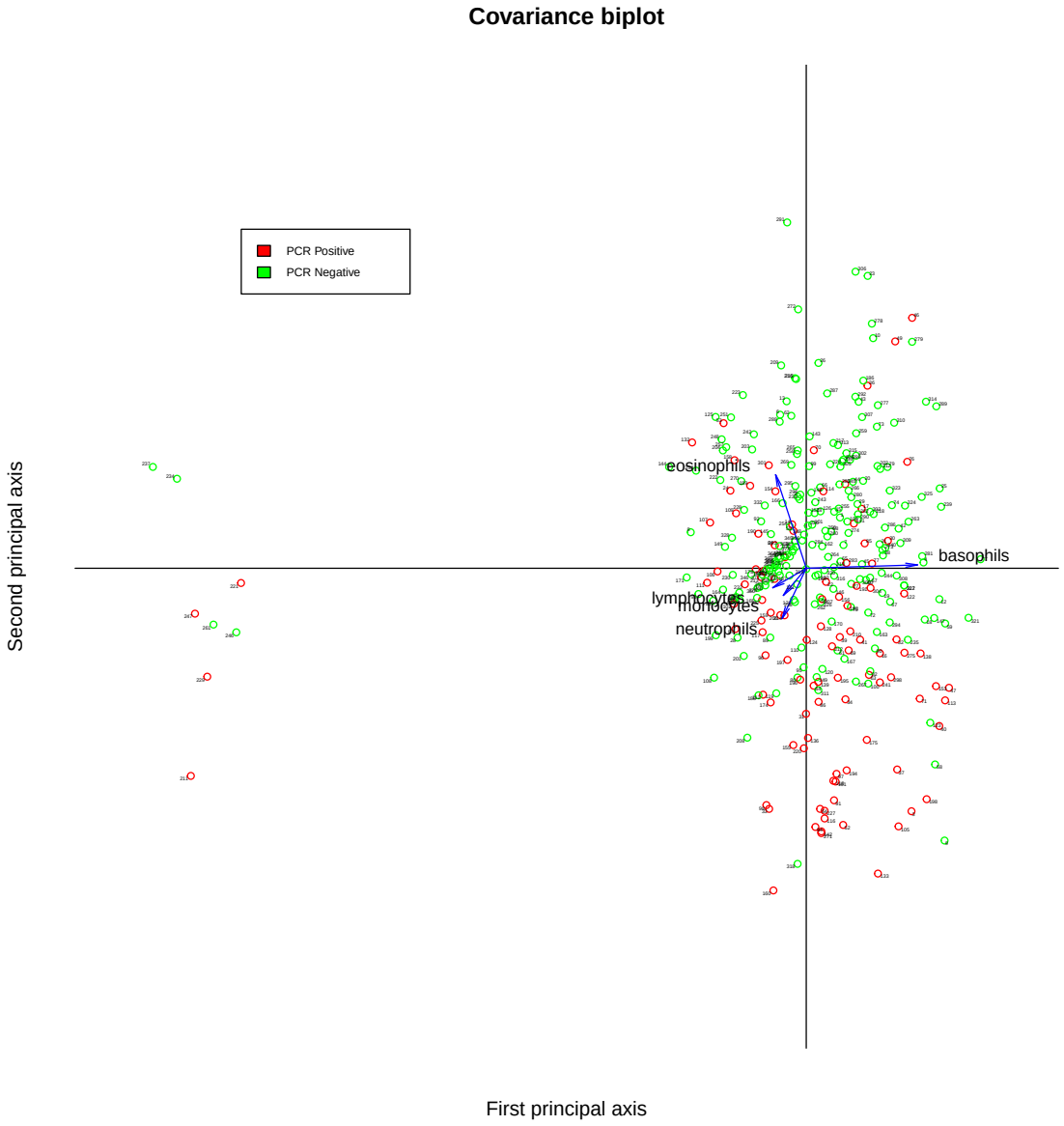

Figure 20: Covariance biplot of PCA using WBC composition data.

From both the form biplot and covariance biplot, we can see that the log ratio of basophils and lymphocytes almost coincides with the first principal axis, and the log ratio of eosinophils and neutrophils almost coincides with the second principal axis. Moreover, the log ratio of eosinophils and neutrophils is almost uncorrelated with the log ratio of basophils and lymphocytes. These aspects are overlooked if we only visualize our dataset by naive PCA and only captured if we view our dataset as a compositional dataset.

The following two biplots are the form biplots and covariance biplots of the WBC compositional dataset if we remove the eight basophils count zero cases. Notice that the direction of basophils and eosinophils vectors changes a lot if we remove the effect of those outliers. However, as we can see in the next section, removal of the outliers hardly affects the classification result as the change hardly affects the LR-LDA vector's coefficients. Therefore, we only include those biplots for reference purpose and still use the processed WBC compositional dataset (329 cases) that includes the eight outliers.

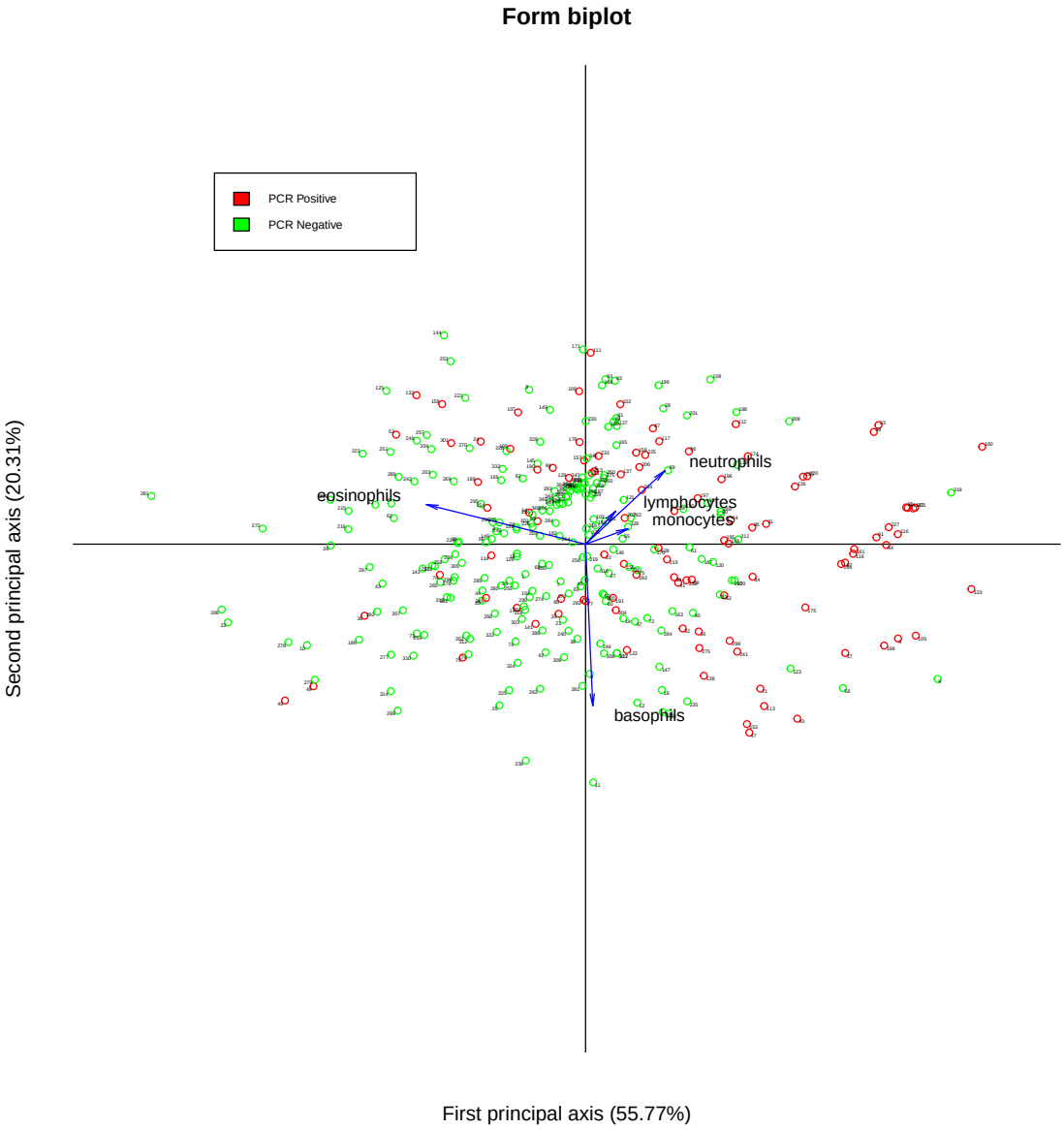

Figure 21: Form biplot of PCA using WBC composition data after removing eight outliers.

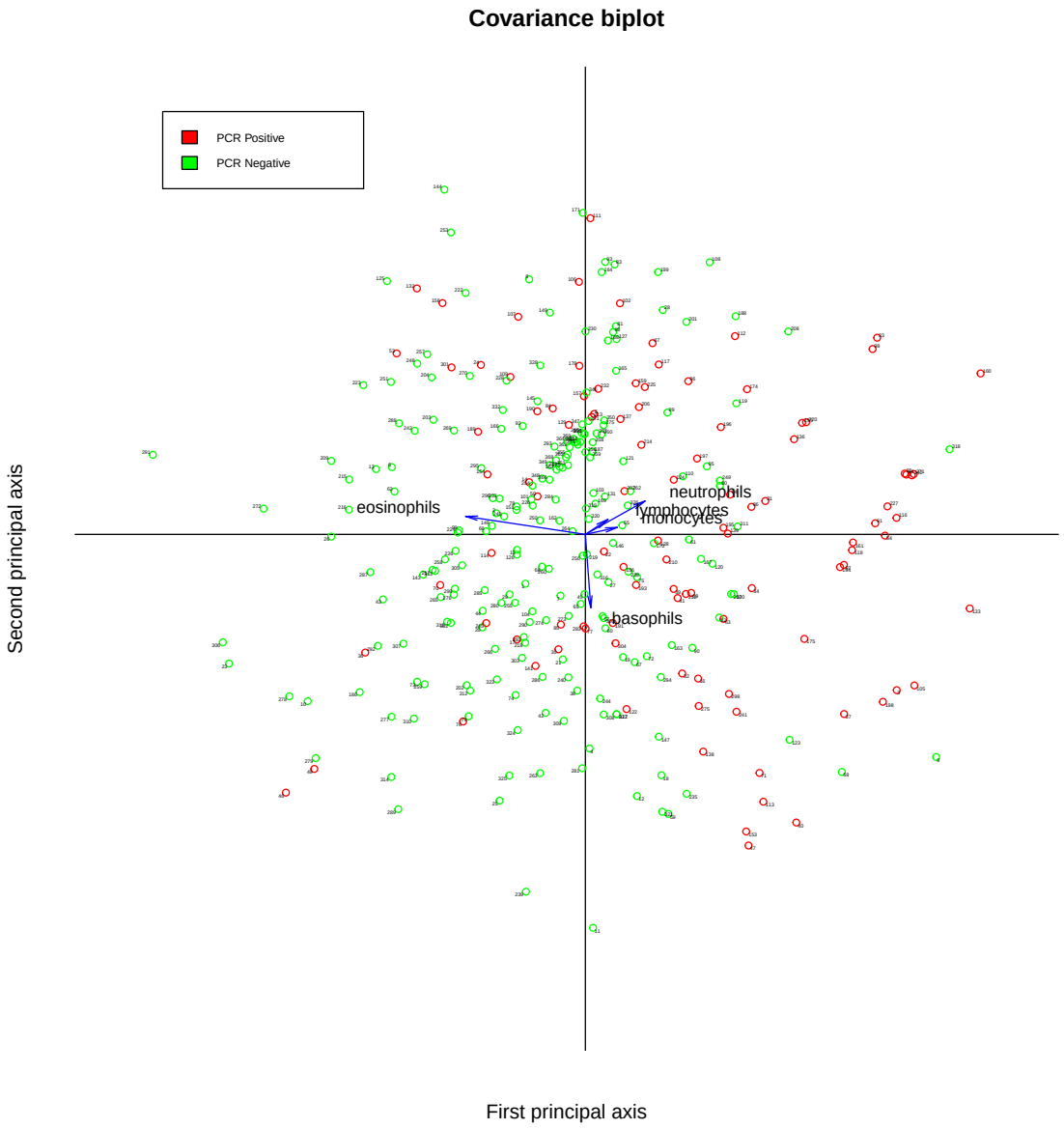

Figure 22: Covariance biplot of PCA using WBC composition data after removing eight outliers.

4.4 LR-LDA

We apply the LR-LDA method by applying the function LRLDA in the R package TOOLSFORCODA [6] to our compositional dataset, and we obtain Figure 23.

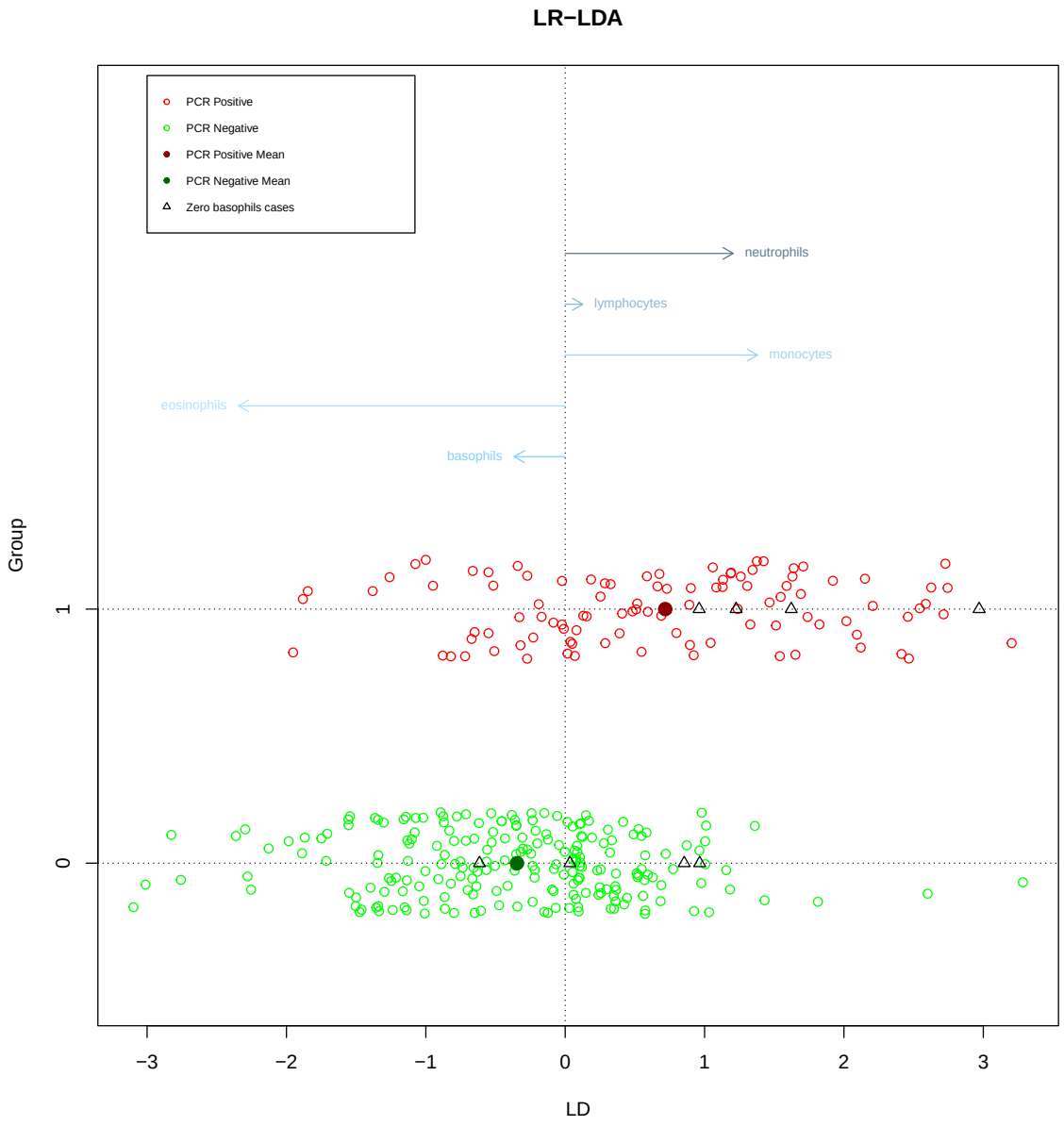

Figure 23: LR-LDA of WBC compositional data

We still add a small random noise (through the "jitter" function in R) to the COVID condition variable to avoid overlapping. The arrows on the top of Figure 23 represent the coefficients obtained in the LR-LDA, it is one dimensional since we are only dealing with two groups. Also, we represent the eight cases with zero basophils as triangles and noticed that they are no longer outliers in the LR-LDA analysis. Similar to the linear coefficient of the LDA analysis part, larger distance between vector arrows can be interpreted as more powerful log ratios when predicting the COVID-19 condition outcome. Notice that the LR-LDA technique performs a little worse in the WBC composition data compared to the WBC counts dataset as we have more misclassified cases. This fact can be seen in the confusion matrix of LR-LDA in Table 11. Moreover, we obtain a classification rate of 77.51% using LR-LDA. A visualization of LR-LDA vectors (coefficients) can be seen in Figure

23 and Table 12.

	Prediction	
True	0	1
0	205	17
1	57	50

Table 11: Confusion matrix for LR-LDA with WBC composition data.

neutrophils	0.40175892
lymphocytes	0.04193869
monocytes	0.45973197
eosinophils	-0.78140020
basophils	-0.12202938

Table 12: Coefficient vector of the LR-LDA method.

From the coefficients vectors in Figure 23 and Table 12, we can see that the log-ratio of neutrophils and eosinophils is one of the most discriminating log-ratio between variables (only slightly less than the log ratio of monocytes and eosinophils), as they have a relative large distance between the arrows $(0.40 - (-0.78) = 1.18)$. This observation also coincides with our LR-PCA biplots (both the form biplot and the covariance biplot), as the log-ratio of neutrophils and eosinophils almost coincide with the second principal axis which has a relative large classification power of PCR positive cases and PCR negative cases. Moreover, the contrast of neutrophils and eosinophils also appeared in the Figure 11, where their difference almost coincides with the second principal axis and explains most of the data structure together with the total WBC counts (which coincides with the first principal axis).

The ROC curve for the LR-LDA method is shown in the ROC Figure 24. The area under the curve (AUC) is 0.7625.

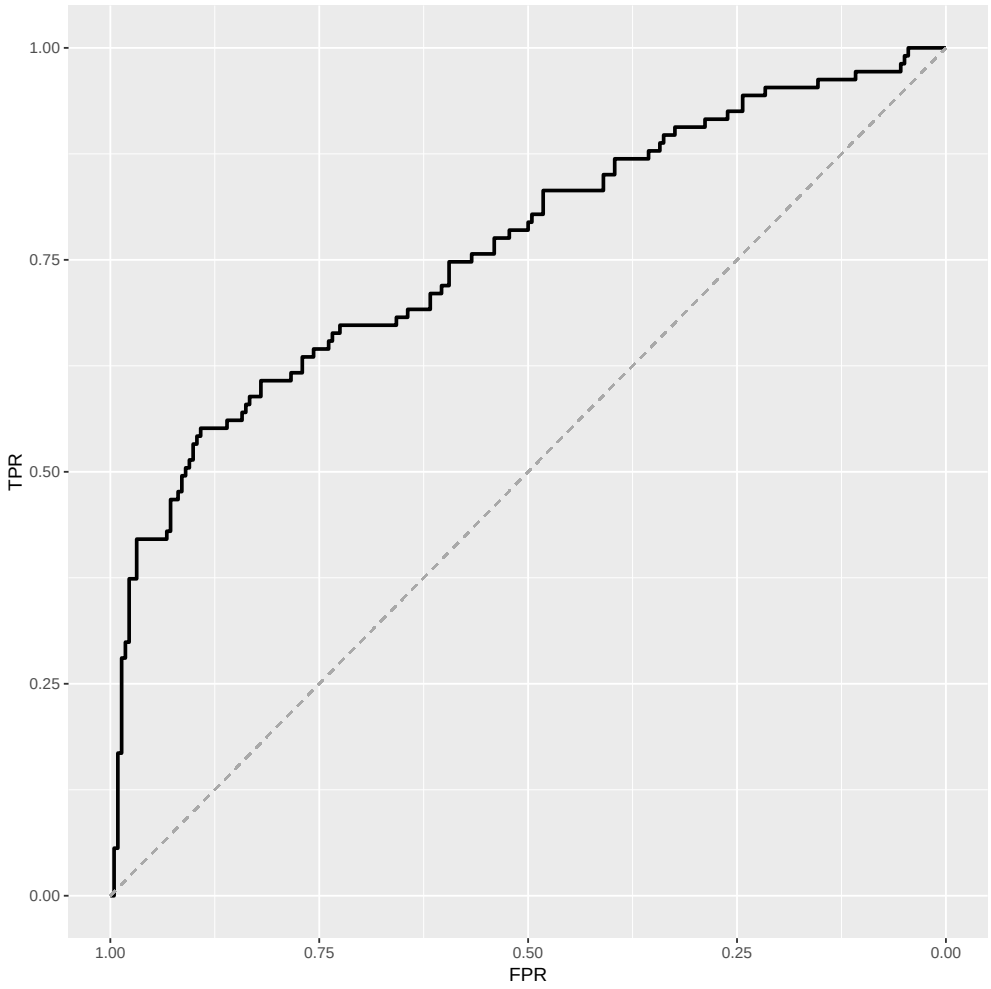

Figure 24: The ROC curve for the LR-LDA method.

If we remove those eight outliers instead, we have the following confusion matrix and coefficient vector of the LR-LDA. Now the classification accuracy is 77.57% instead of 77.51%. We can see the difference of the classification results is small by comparing Table 11 with Table 13 and comparing coefficient vector Table 12 with coefficient vector Table 14. As we mentioned by the end of LR-PCA section, removing eight outliers hardly affects the classification result and the change doesn't affect much of LR-LDA vector's coefficient. Therefore, we keep the eight outliers in the analysis.

	Prediction	
True	0	1
0	203	15
1	57	46

Table 13: Confusion matrix for LR-LDA with WBC composition data removing eight outliers.

neutrophils	0.37004853
lymphocytes	0.06512380
monocytes	0.44244536
eosinophils	-0.81160817
basophils	-0.06600953

Table 14: Coefficient vector of the LR-LDA method removing eight outliers.

4.5 Feature selection

By applying LR-PCA and LR-LDA to the WBC compositional data did give us some extra insight of its structure, but they don't give us a satisfactory classification rate that improves over the rate obtained with the full dataset LDA using all information. Therefore, we try to add some extra variables to the compositional dataset hoping to achieve a better classification rate. Notice that just adding the clr transformed cell counts to the full dataset will cause us singularity problem as the clr transformed counts are systematically collinear. Therefore, we apply the ilr-transformation as in Equation (36) (instead of clr as in Equation (34)) to our WBC counts dataset to avoid the collinearity problem in the whole feature selection section. Now our WBC compositional dataset is represented by four log-ratios. We note that without additional non-compositional predictors, the classification results and AUC scores will be exactly the same using ilr-transformation and clr-transformation, since both methods give the exact same posterior probabilities.

4.5.1 Adding total WBC counts

One thing that was sometimes overlooked in the compositional data analysis is the total number of all the components. Although in CoDA we only care about the relative values in our dataset, their total number, which is the total number of white blood cell counts can be a powerful predictor for the outcome COVID-19 infection status. Therefore, we calculate the summation of WBC counts of each patient and treat it as a new variable. Then we include this total WBC counts variable into our WBC compositional dataset and apply LR-LDA method. Since now we don't have the singularity covariance matrix problem as we choose to use ilr-transformation for feature selection, specific programs that use generalized inverses are no longer needed and the analysis can be performed with the standard R function LDA.

The result for the LR-LDA classification using ilr-transformed WBC composition data plus the total WBC counts variable are presented in the Confusion Matrix below (see Table 15). The classification accuracy is 79.64% before cross-validation. This result is slightly better than the LR-LDA classification itself but still not as good as the LDA classification using the full dataset. The ROC curve of using ilr-transformed WBC composition data plus the total WBC counts variable is presented in the Figure 25. This analysis shows that complementing the log-ratio transformed compositions with their total count can be helpful.

	Prediction	
True	0	1
0	202	47
1	20	60

Table 15: Confusion matrix for the LR-LDA with WBC composition data adding total WBC counts.

4.5.2 Step-wise LDA

Another common approach for feature selection is to use a step-wise algorithm. We use the TRAIN function with "METHOD = STEPLDA" from the R package CARET to apply step-wise LDA automatically [10]. In the TRAIN function, a grid of all parameters is created and the model is trained on slightly different data for each candidate combination of tuning parameters. Across each data set, the performance of held-out samples is calculated and the mean and standard deviation is summarized for each combination. The combination with the optimal resampling statistic (in our case this is the classification accuracy) is chosen as the final model and the entire training set is used to fit a final model[10]. When we fit the Step-wise LDA model, we choose from the full dataset variables excluding WBC counts variables ($15 - 5 = 10$ of them, see Table 5) plus the ilr transformed compositional variables (4 of them) and the total WBC counts variable. And the outcome is the COVID-19 condition variable. With a 10-fold cross validation while selecting variables, the TRAIN function give us the final model as:

$$Condition \sim platelets. \tag{67}$$

The classification accuracy for this model is 73.56% before cross-validation. This result is actually worse compared to the LR-LDA classification and the LDA with WBC composition data adding total WBC counts. The confusion matrix is presented in Table 16 below. The ROC curve of using step-wise LDA model is presented in the Figure 25.

	Prediction	
True	0	1
0	209	74
1	13	33

Table 16: Confusion matrix for the LDA with the step-wise chosen model.

Also, we can fit the Step-wise LDA model over the full dataset without adding compositional variables and the total WBC counts variable, to make our result comparable to the full dataset LDA. The TRAIN function give us the final model as:

$$Condition \sim lymphocytes + hematocritP. \tag{68}$$

The classification accuracy for this model is 80.51% before cross-validation, which is worse compared to the naive LDA with the full dataset but gives a more parsimonious model. Its confusion matrix is presented in Table 17 below and its ROC curve is presented in the Figure 25.

	Prediction	
True	0	1
0	209	51
1	13	56

Table 17: Confusion matrix for the LDA with the step-wise chosen model on the full dataset.

4.5.3 Adding predictive variables

Even the classification accuracy of the first model in this section (compositional dataset plus total WBC counts model) improved the naive LR-LDA compositional data model, it still fails to improve

over the full LDA model which has a classification rate of 83.587% before cross-validation. However, the full LDA model has given us some insight of each variables' prediction power, and we try to exploit this information by adding the most predictive variables other than WBC counts (i.e. variables with a large linear discriminant coefficient in absolute value in Table 5) and the total WBC counts variable to the compositional dataset and apply the LR-LDA method. The reason we do not add raw WBC counts variables is that their information should be fully incorporated in the compositional dataset as we view WBC counts as compositional variables and only care about their relative values.

As a result, when we add the red blood cells, the platelets and the total WBC counts variable to the ilr-transformed dataset and apply the LDA method to those seven variables, we obtain a classification accuracy rate of 85.11% before cross validation and a classification accuracy rate of 84.50% after cross validation. Both results are actually better compared to the best result we had before by using LDA on the full dataset (which has a classification rate of 83.58% before cross-validation and 81.46% after cross-validation). The before cross-validation confusion matrix is presented below in Table 18 and the after cross-validation confusion matrix is presented below in Table 19. The ROC curve is presented in Figure 25.

	Prediction	
True	0	1
0	207	34
1	15	73

Table 18: Confusion matrix for the LDA with compositional data adding predictive variables (before CV).

	Prediction	
True	0	1
0	206	35
1	16	72

Table 19: Confusion matrix for the LDA with compositional data adding predictive variables (after CV).

One thing to notice is that without the four compositional variables (given by ilr-transformation), simply applying the LDA method to the three added variables (red blood cells, platelets, and total WBC counts) will actually give us much worse results: classification accuracy of 76.60% before cross-validation and 75.08% after cross-validation. Also, if we just use the non-compositional LDA model with WBC counts variables adding predictive variables (red blood cells and platelets, no total WBC counts variables to avoid collinearity problem), we will also obtain worse results: classification accuracy of 82.98% before cross-validation and 82.07% after cross-validation. All of the results presented above are worse compared to the current compositional model, which means adding compositional variables to our LDA model does help us get a better understanding of data structure and help us improve classification rate.

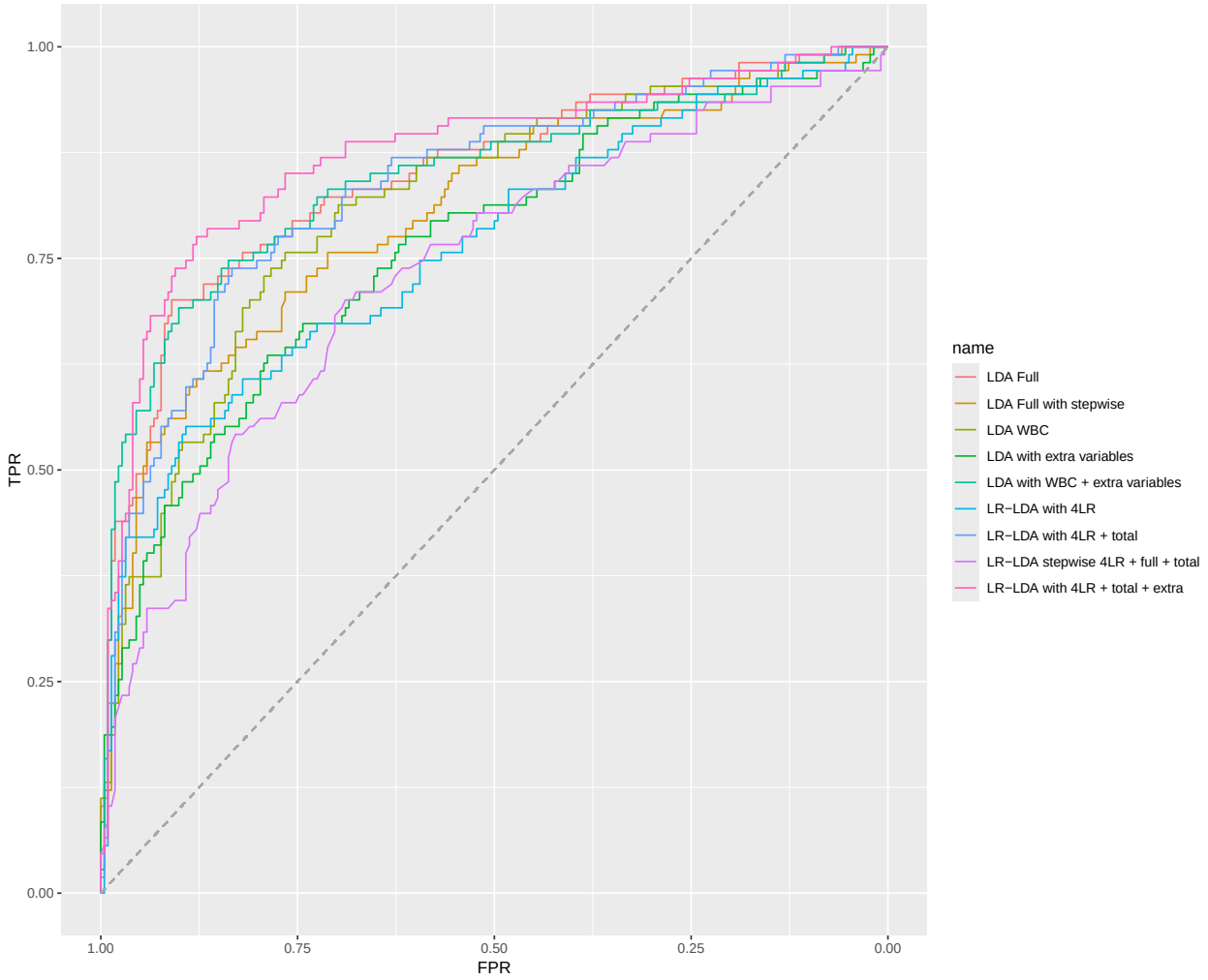

Figure 25: The ROC curves for all models.

To highlight the ROC curves for compositional LDA models, a separate Figure 26 only including these ROC curves is presented below.

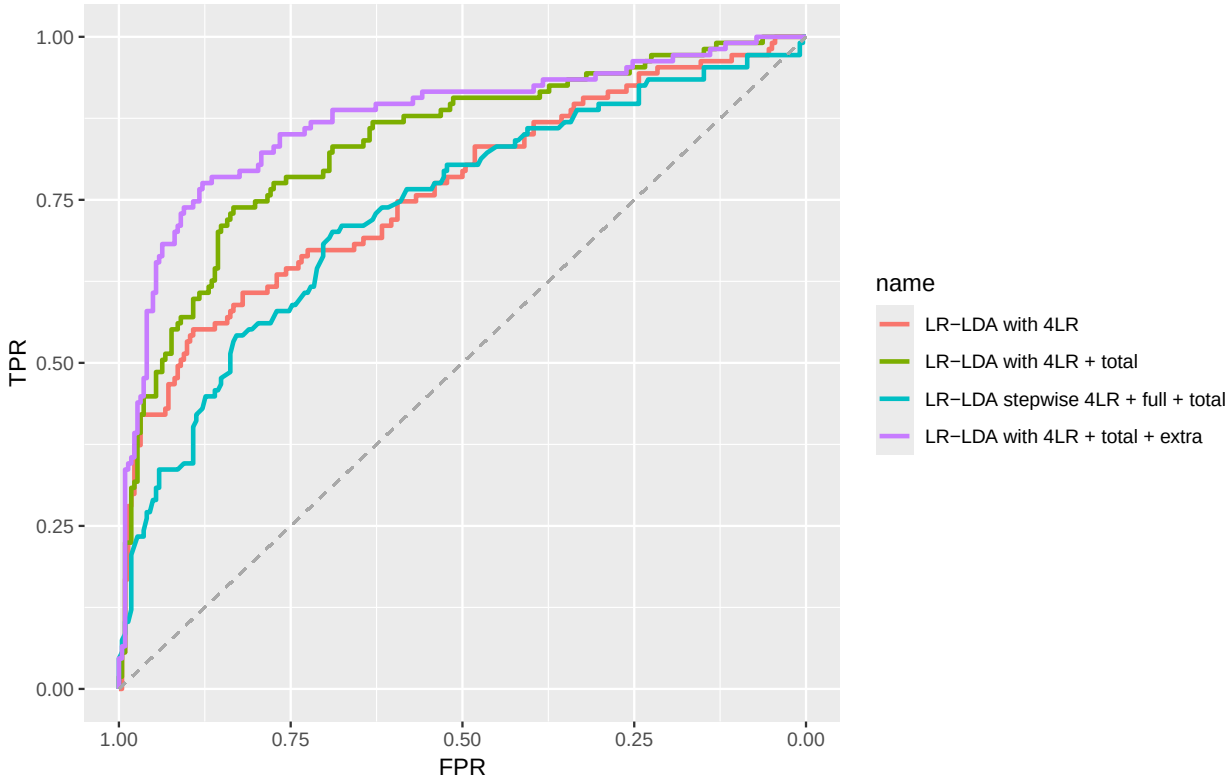

Figure 26: The ROC curves for compositional models.

Table 20 and Table 21 that summarize our models and their performances so far are presented below.

Non-compositional Models	Accuracy (before CV)	AUC (before CV)	Accuracy (after CV)
LDA Full	83.59%	0.8487	81.46%
LDA Full with step-wise	80.55%	0.8028	80.55%
LDA WBC	77.51%	0.8129	76.90%
LDA with extra variables	76.60%	0.7649	75.08%
LDA with WBC + extra variables	82.98%	0.8456	82.07%

Table 20: The table of all non-compositional models and their performances.

Compositional Models	Accuracy (before CV)	AUC (before CV)	Accuracy (after CV)
LR-LDA with 4LR	77.51%	0.7625	76.90%
LR-LDA with 4LR + total	79.64%	0.8354	79.03%
LR-LDA stepwise 4LR + full + total	73.56%	0.7384	73.56%
LR-LDA 4LR + Total + extra	**85.11%**	**0.8727**	**84.50%**

Table 21: The table of all compositional models and their performances.

5 Discussion and Conclusions

In this section, we discuss some issues we have encountered when doing data analysis and summarize the main findings of this thesis.

5.1 Low sensitivity issue about the PCR test

In our dataset, about 67% of the clinical cases is PCR-negative in our cleaned dataset, whereas all of them were symptomatic before taking the blood test and PCR test. According to Prof. Márcio Dorn, one of the original data generators and maintainer of the Ecuador CBC database [4], the situation in which people with symptoms show negative PCR results is complex and can have several explanations.

We can think about a few scenarios that may have contributed to these results. The first reason for false negatives in PCR is related to the infection stage of patients and the different variants of the virus. The PCR test is most sensitive during the period of highest viral load, which usually occurs in the first few days after the onset of symptoms. If the tests were carried out later in the infection, the viral load may have decreased enough to generate a false negative. On the contrary, some patients who have been tested positive may be at the end of the immunological window while others at the beginning. In other words, the binary COVID-19 infection status variable is not enough to capture the continuous infection stage information, which make our classification attempt more complex and difficult. Also, the constant emergence of new variants of the virus during the pandemic, with their mutations, may have affected the sensitivity of the PCR test, leading to an increase in false negatives.

The presence of other pathogens, such as other respiratory viruses, and allergies can also lead to negative cases in the PCR, and those cases are not necessarily *false* negatives. For example, the symptoms may result from other common respiratory viruses, such as influenza or rhinovirus, circulating simultaneously with SARS-CoV-2. According to Prof. Dorn, this is the most likely cases for most negative PCR results in the Ecuador database. Also, in some cases, symptoms such as runny nose, cough, and sore throat can be triggered by seasonal or other allergies, and not necessarily by a viral infection. Consequently, part of the presumably negative cases are, despite being symptomatic, in fact truly PCR-negative for they do not suffer from COVID-19, but from a different respiratory health problem.

There are also other factors that can contribute to this phenomenon. For example, The perception of symptoms can vary from person to person, and some individuals may report symptoms even in the absence of an infection. Also, the immune system can control the infection quickly in some cases, resulting in a low viral load and a negative PCR test, even in the presence of symptoms.

Unfortunately, there is no single right answer. There are many possibilities and we need further investigation if we try to find out reasons behind both true and false negative cases.

5.2 Conclusions

The data analysis shows that PCA and LR-PCA can be useful for exploring the data. In particular, biplots are shown to be powerful for visualization, both compositional, non-compositional, and mixed data. LDA and LR-LDA are shown to be efficient classification tools for categorical outcomes,

and we are able to quantify the predictive power of each variable which turns out to be helpful for feature selection and potential interpretation. All computations and figure generation were completed using the statistical environment R [18], and the computational load was manageable on a laptop computer with 12th Gen Intel(R) Core(TM) i7-12700H 2.70 GHz processor and 16 Gigabytes RAM.

Moreover, we have seen that compositional data analysis methods (LR-PCA and LR-LDA) have given us some extra insights about the data structure and the best classification model for COVID-19 status prediction.

To be more specific, from the LR-PCA visualization we have known that the log-ratio of neutrophils and eosinophils is a key feature for infection status classification. This information can not be obtained by the PCA with the full dataset or the PCA with the WBC counts dataset. And from the LR-LDA coefficients, we noticed that the log-ratio of neutrophils and eosinophils has relatively large classification power which confirms our observations again.

By comparing the classification accuracy and the AUC of all models, we have shown that the compositional model with some extra variables has the best performance. This performance is due to the extra information we gained by adding compositional variables, as the model with only added variables or the model with the non-compositional WBC counts with extra variables performed worse compared to our compositional model.

To summarize, we have successfully illustrated the efficacy of combining compositional data analysis with traditional non-compositional data-visualization method and classification models. When our dataset has an intrinsic and scientifically meaningful compositional structure (like the WBC counts variables), we recommend combining compositional data analysis method with the classic analysis methods (like LR-LDA and LR-PCA) to gain extra insights and improve the performance of models. Moreover, we have successfully come up with a compositional classification model using only complete blood-count data. This compositional model satisfied our expectation to be a cheap and efficient tool to help diagnosing COVID-19 infection with a superior cross-validated accuracy.

5.3 Future extensions

Although in this thesis we have been focusing on LDA and LR-LDA for classification models, other common methods (like logistic regression, K-nearest neighbor algorithm, Quadratic discriminant analysis, etc) could also be applied with compositional data and evaluated. Also, there are more advanced classification models like neural networks, support vector machines, LIME [19] and SHAP [13], and the inclusion of compositional information in the form of log-ratio transformed compositions using these methodologies is another interesting topic for future investigation.

The discussion above clearly suggests that the PCR-negative cases form a heterogeneous group, and this suggests a potential direction for future research. As Prof. Dorn mentioned, the situation of symptomatic patients showing negative PCR results is complex and has multiple explanations that are quite different from each other. We suggest the use of cluster analysis to potentially discover groupings in the negative cases. This may make our compositional classification model more accurate and interpretable.

Another interesting extension would be redoing the analysis and the model building process with a larger and more comprehensive dataset. Incorporating additional baseline variables, such as detailed clinical data, pre-existing medical conditions, and blood oxygenation levels, may significantly enhance the predictive power and the interpretation of the model.

6 References

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. The Blackburn press, Caldwell, NJ, 1986. 2003 printing.

[2] J. Aitchison and M Greenacre. Biplots of compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4):375–392, 2002.

[3] D. Dutta, S. Naiyer, S. Mansuri, N. Soni, V. Singh, K.H. Bhat, N. Singh, G. Arora, and M S. Mansuri. Covid-19 diagnosis: a comprehensive review of the rt-qpcr method for detection of sars-cov-2. *Diagnostics*, 12(6):1503, 2022.

[4] B.C. Feltes, I. Vieira, J. Parraga-Alava, J. Meza, E. Portmann, L. Terán, and M. Dorn. Feature selection reveal peripheral blood parameter's changes between covid-19 infections patients from brazil and ecuador. *Infection, Genetics and Evolution*, 98:105228, 2022.

[5] K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[6] J. Graffelman, V. Pawlowsky-Glahn, J. Jose Egozcue, and A. Buccianti. Exploration of geochemical data with compositional canonical biplots. *Journal of Geochemical Exploration*, 194:120–33, 2018.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014.

[8] R.A. Johnson and W.W. Dean. *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2006.

[9] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer New York, 2013.

[10] M. Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[11] M. Leach, H. MacGregor, I. Scoones, and A. Wilkinson. Post-pandemic transformations: How and why covid-19 requires us to rethink development. *World development*, 138:105233, 2021.

[12] A.T. Levin, N. Owusu-Boaitey, S. Pugh, B. Fosdick, A.B. Zwi, A. Malani, S. Soman, L. Besançon, I. Kashnitsky, S. Ganesh, et al. Assessing the burden of covid-19 in developing countries: systematic review, meta-analysis and public policy implications. *BMJ Global Health*, 7(5):e008477, 2022.

[13] S. M Lundberg and S. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[14] J. Palarea-Albaladejo and J. A. Martín-Fernández. zCompositions – R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

[15] W. E. Paul. *Fundamental immunology*. Lippincott Williams & Wilkins, 2012.

[16] V. Pawlowsky-Glahn, J.J. Egozcue, and R. Tolosana-Delgado. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.

[17] R. Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.

[18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[20] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

7 Appendix

All of the R code and dataset used for this Master thesis can be found in `https://github.com/ZhilongZ2/MS-Thesis-code`.