



# **A NEW RNA-SEQ DATA CLASSIFIER BASED ON QUANTILE TRANSFORMATION**

**NECLA KOÇHAN**

Ph.D. Thesis

Graduate School  
Izmir University of Economics

Izmir

2020

**A NEW RNA-SEQ DATA CLASSIFIER BASED ON  
QUANTILE TRANSFORMATION**



**NECLA KOÇHAN**

A Thesis Submitted to  
The Graduate School of Izmir University of Economics  
Applied Mathematics and Statistics Program in Mathematics

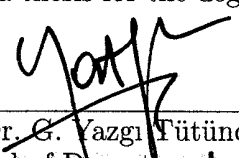
Izmir  
2020

**Ph.D. DISSERTATION EXAMINATION RESULT FORM**

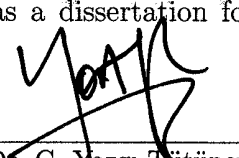
Approval of the Graduate School

  
Prof. Dr. Mehmet Efe Biresselioglu  
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy.

  
Prof. Dr. G. Yazgi Tutuncu  
Head of Department

We have read the dissertation entitled “**A new RNA-Seq Data Classifier based on quantile transformation**” completed by NECLA KOÇHAN under supervision of **Prof. Dr. G. Yazgi Tutuncu** and we certify that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

  
Prof. Dr. G. Yazgi Tutuncu  
Supervisor

**Examining Committee Members**

Date: 16/01/2020

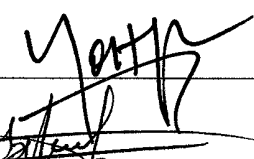
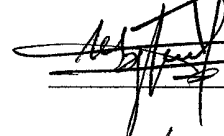

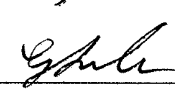

Prof. Dr. G. Yazgi Tutuncu  
Dept. of Mathematics, IUE

Prof. Dr. İsmihan Bayramoglu  
Dept. of Mathematics, IUE

Prof. Dr. Burcu Hüdaverdi Üçer  
Dept. of Statistics, Dokuz Eylül University

Assoc. Prof. Dr. Güvenç Arslan  
Dept. of Statistics, Kırıkkale University

Assoc. Prof. Dr. Zeynep Fırtına Karagonlar  
Dept. of Genetics and Bioengineering, IUE

# ABSTRACT

## A NEW RNA-SEQ DATA CLASSIFIER BASED ON QUANTILE TRANSFORMATION

Koçhan, Necla

PhD in Applied Mathematics and Statistics

Advisor: Prof. Dr. G. Yazgı Tütüncü

January, 2020

Recently in cancer research, true classification of the sub-type of a patient with a particular cancer, leads a better predictive and a customized treatment for that patient. Therefore, classification of a patient to a cancer sub-type has a crucial importance and can be done by using genetic information. Most of the existing classifiers assume that genes are independent; however, this is not a realistic approach for real RNA-Seq classification problems. For this reason, in this thesis a new classifier, which incorporates the dependence structure between genes into a model, is proposed. The dependency between genes is first modelled by sample covariance matrix and then by local covariance matrix. The local covariance matrix is estimated by the local dependency approximation. The classification algorithm is coded in R programming language and a new classification package for RNA-Seq data is developed. The performance of this new classifier is compared with the existing classifiers over real RNA-Seq data sets, in terms of classification error rates.

Keywords: Quadratic discriminant analysis, RNA-Sequencing, gene expression, dependence, local covariance matrix, regularization, classification.



# ÖZET

## KUANTİL TRANSFORMASYON TABANLI YENİ BİR RNA- SEKANS VERİ SINIFLANDIRICISI

Koçhan, Necla

Uygulamalı Matematik ve İstatistik Doktora Programı

Advisor: Prof. Dr. G. Yazgı Tütüncü

Ocak, 2020

Son zamanlarda kanser arařtırmalarında, bilinen bir kanser tipi olan bir hastanın o kanserin çeşidine göre doğru sınıflandırılması o hasta için daha iyi tahminlere dayanan ve kişiye özel tedavi sağlamaktadır. Bu nedenle, hastanın kanser çeşidine göre sınıflandırılması çok önemlidir ve bu, genetik bilgi kullanılarak yapılabilmektedir. Mevcut sınıflandırıcıların çoğu genlerin bağımsız olduğu varsayımına dayanmaktadır; ancak, bu varsayım asıl RNA-Sekans sınıflandırma problemleri için gerçekçi bir yaklaşım değildir. Bu nedenle, bu tezde, genler arasındaki bağımlılık yapısını dikkate alan yeni bir sınıflandırıcı önerilmektedir. Genler arasındaki bağımlılık önce kovaryans matrisi ve daha sonra lokal kovaryans matrisi ile modellenmektedir. Lokal kovaryans matrisi, lokal bağımlılık fonksiyonu kullanılarak tahmin edilmektedir. Sınıflama algoritması R programlama dilinde kodlanmış olup RNA-Sekans verileri için yeni bir sınıflama paketi geliştirilmiştir. Yeni sınıflandırıcının performansı, gerçek RNA-Sekans verileri kullanılarak mevcut sınıflandırıcılar ile sınıflandırma hataları açısından

karşılaştırılmıştır.

Anahtar Kelimeler: Karesel diskriminant analizi, RNA-Sekanslama, gen ifadesi, bağımlılık, lokal kovaryans matrisi, sınıflama.



Dedicated to my family and my beloved mother...





## ACKNOWLEDGEMENT

I would like to start by giving my gratitude to my supervisor Prof. Dr. G. Yazgı Tütüncü. She has been always beside me, helping and guiding, from the moment I started my academic career. She has always respected my ideas, supported me for what I wish to study and helping me throughout all the process. Her insightful remarks and suggestions have helped me to think through and overcome my problems. Moreover, I want to thank her for the days she spent on reading and commenting on the thesis.

I would like to thank Prof. Dr. İsmihan Bayramoğlu and Assoc. Prof. Dr. Güvenç Arslan for dedicating their valuable time to guide me throughout my study. Their ideas and studies have always been a great example for me.

I would also like to extend my deepest gratitude to my anchor, Dr. Göknur Giner who has supported me since the very beginning of my thesis, helped me to study in Walter and Eliza Hall Institute Bioinformatics Division and to enhance the breadth of my knowledge.

I'm deeply indebted to Dr. Luke Gandolfo who supported me in many ways including the development of the model used in this thesis and the publication stages of our article, and provided me with the moral support I needed to pull through whenever I needed.

I would like to extend my sincere thanks to Prof. Dr. Gordon Smyth who invited me to his lab and supported me both financially and morally.

I would also like to thank to Prof. Dr. Terry Speed for helping me during my visit at Walter and Eliza Hall Institute.

I would like to give my special thanks to Walter and Eliza Hall Institute, Bioinformatics division and Izmir University of Economics, Department of Mathematics for giving me chances to use all the resources I need to complete

my thesis.

I sincerely thank my colleagues at WEHI, many friends I met during this journey and my housemate Gabrielle Baker for many great memories. I am much grateful to spend the most wonderful time of my life with them.

I also thank to my colleagues at Izmir University of Economics who were very understanding and supportive throughout this process. I would especially like to thank Özge Altıntaş, Dr. Aslı Güldürdek and Dr. Cihangir Kan for listening to me for hours, for their great friendship and for their timeless support in every aspect of my life.

My heartfelt gratitude is to my spouse Cemal Koçhan, he has been beside me from the beginning of the process, listening to my complaints, giving me advice and most importantly believing in me whenever I doubt myself. I also thank to my big family for their unconditional support throughout my life. I couldn't have been able to finish this journey without their constant encouragement and support.

Last, I would like to thank The Scientific and Technological Research Council of Turkey (TUBITAK) for the scholarship they gave me to help me to complete the PhD program. I hope my thesis will help and guide many researchers.

## TABLE OF CONTENTS

ABSTRACT .....	iii
ÖZET .....	v
ACKNOWLEDGEMENT .....	viii
TABLE OF CONTENTS .....	x
LIST OF TABLES .....	xii
LIST OF FIGURES .....	xiii
LIST OF ABBREVIATIONS .....	xiv
CHAPTER 1: INTRODUCTION .....	1
CHAPTER 2: RNA SEQUENCING AND DISCRIMINANT ANALYSIS .....	5
2.1 <i>Basic Concepts and Terms</i> .....	5
2.2 <i>Next Generation RNA Sequencing</i> .....	7
2.3 <i>RNA-Seq Data</i> .....	9
2.4 <i>Notations</i> .....	10
2.5 <i>Discrete Models Proposed for RNA-Seq Data</i> .....	11
2.5.1 <i>Poisson Model</i> .....	11
2.5.2 <i>Negative Binomial Model</i> .....	12
2.6 <i>Discriminant Analysis</i> .....	13
CHAPTER 3: METHODS FOR RNA-SEQ DATA CLASSIFICATION .....	16
3.1 <i>Machine Learning Methods Applied on RNA-Seq Data</i> .....	16
3.1.1 <i>k-Nearest Neighbour Algorithm</i> .....	16
3.1.2 <i>Support Vector Machines</i> .....	17
3.1.3 <i>Logistic Regression Classifier</i> .....	18
3.2 <i>Discrete Methods Applied on RNA-Seq Data</i> .....	20
3.2.1 <i>Poisson Linear Discriminant Analysis</i> .....	20
3.2.2 <i>Negative-Binomial Linear Discriminant Analysis</i> .....	21
3.2.3 <i>Voom Based Diagonal Linear Discriminant Analysis</i> .....	22
CHAPTER 4: THE MODEL .....	24
4.1 <i>General Structure of The Model</i> .....	24
4.2 <i>qtQDA Model</i> .....	25
4.2.1 <i>Preprocess</i> .....	25
4.2.2 <i>Gene Selection</i> .....	26
4.2.3 <i>Parameter Estimation I</i> .....	26

4.2.4 Quantile Transformation .....	28
4.2.5 Parameter Estimation II .....	29
4.2.6 Classification .....	30
4.3 <i>Application on Real RNA-Seq Data</i> .....	32
4.3.1 Experimental Data Sets .....	32
4.3.2 Implimentation of Existing Classifiers .....	33
4.3.3 Evaluation of the Performance of Claccifiers .....	34
4.4 <i>Results</i> .....	35
4.5 <i>Discussion</i> .....	37
<b>CHAPTER 5: LOCAL COVARIANCE MATRIX ESTIMATION .....</b>	<b>39</b>
5.1 <i>Local Dependence Functions for Multivariate Normal Distributions</i> ....	40
5.1.1 The Local Dependence Function .....	40
5.2 <i>A New Estimate of Local Dependence Function</i> .....	43
5.3 <i>Results</i> .....	43
5.4 <i>Discussion</i> .....	44
<b>CHAPTER 6: CONCLUSION AND FURTHER STUDIES .....</b>	<b>47</b>
<b>REFERENCES .....</b>	<b>49</b>

## LIST OF TABLES

Table 2.1. RNA-Seq data matrix of cervical cancer data .....	10
Table 4.1. Confusion matrix .....	34
Table 4.2. Minimum error rate achieved for each classifier in each data set .....	36
Table 5.1. Error rates for cervical cancer and HapMap data sets.....	44



## LIST OF FIGURES

Figure 2.1. The detailed structure of DNA and RNA .....	6
Figure 2.2. Transcription and translation processes .....	7
Figure 2.3. Technical replicates vs biological replicates .....	7
Figure 2.4. A workflow of RNA Sequencing .....	8
Figure 2.5. Gene Expression Data Matrix .....	9
Figure 3.1. An example of kNN algorithm when $k=3$ .....	17
Figure 3.2. Kernel method .....	18
Figure 4.1. Classification error rate as a function of the number of genes chosen for classification of the (a) Cervical Cancer, (b) Prostat Cancer and (c) HapMap data sets .....	36
Figure 5.1. Classification error rate as a function of the number of genes chosen for classification of the cervical cancer .....	45
Figure 5.2. Classification error rate as a function of the number of genes chosen for classification of the HapMap data .....	45

## LIST OF ABBREVIATIONS

APL: Adjusted Profile Likelihood

cDNA: complementary Deoxyriboucleic Acid

CER: Classification Error Rate

DE: Differentially Expressed

DLDA: Diagonal Linear Discriminant Analysis

DNA: Deoxyriboucleic Acid

FN: False Negative

FP: False Positive

GLMnet: Lasso and Elastic-Net Regularized Generalized Linear Models

GLMs: Generalized Linear Models

$k$ NN:  $k$ -Nearest Neighbor

LDA: Linear Discriminant Analysis

LRT: Likelihood Ratio Test

L-qtQDA: Local quantile transformation Quadratic Discriminant Analysis

miRNA: micro Ribonucleic Asid

MLE: Maximum Likelihood Estimator

mRNA: messenger Ribonucleic Asid

MVN: Multivariate Normal

NB: Negative Binomial

NBLDA: Negative Binomial Linear Discriminant Analysis

PLDA: Poisson Linear Discriminant Analysis

QDA: Quadratic Discriminant Analysis

qtQDA: quantile transformation Quadratic Discriminant Analysis

RNA: Ribonucleic Asid

rRNA: ribosomal Ribonucleic Asid

SQDA: Sparse Quadratic Discriminant Analysis

SVMs: Support Vector Machines

TN: True Negative

TP: True Positive

tRNA: transfer Ribonucleic Asid

voomDLDA: voom based Diagonal Linear Discriminant Analysis

## CHAPTER 1: INTRODUCTION

Classification on the basis of RNA sequencing data is an important problem of modern personalized medicine, particularly for disease diagnosis and personalized treatment. For instance, breast cancer has several distinct types where some types respond better to certain treatments and some respond worse. Thus, it is important to know what type of breast cancer a patient has so that the right treatment can be assigned. One approach is to perform RNA sequencing on a cancer sample from the patient and use the gene expression profile as data for classifying what type of cancer the patient has. Knowing the gene expression profile inside a cell gives us important insights into biological processes, e.g. the mechanisms of disease or individual's susceptibility to a certain cancer type. To know the mechanisms of the disease allows the treatment to be personalized and increases the survival chance of the patient.

One can measure gene expression profiles using various high-throughput technologies such as microarray and next generation RNA sequencing (RNA-Seq) technologies. Recently, RNA-Seq has become very popular and widely applied approach in molecular biology studies due to the many advantages. Unlike microarray technologies, RNA-Seq, for instance, can discover the novel transcripts. It can also measure the tens of thousands of genes simultaneously which reduces the sequencing cost (Fu et al., 2009; Haas and Zody, 2010). With these advantages, gene expression data are being generated in large output. Thus, an impressive data analysis task is required to efficiently extract significant amount of biological information from the huge and fast-growing gene expression data.

There are two main goals in gene expression data analysis: (1) to identify differentially expressed genes related to the condition (e.g. tumor and non-tumor samples) (2) to develop a classification model for diagnostic purpose. In this thesis, we focus on the second goal. More precisely, in this thesis, our aim is to determine whether a patient has a disease or not (e.g. tumor or non-tumor) or whether a patient has a specific type of a disease (e.g. type of a breast cancer). We note here that we have only considered the case where we have just two



classes, that is we focused on binary classification problems.

In order to construct a successful classification model, which can be applied on RNA-Seq data, powerful statistical methods are required. Those methods should be able to overcome the problems arises from the high dimensional RNA-Seq data, detect the most informative and minimal gene set to be used in classification and to classify patients correctly. Since RNA-Seq counts the number of reads mapped on to genes and measures gene expression levels on discrete scale, many machine learning algorithms cannot be directly applied to RNA-Seq data. Thus, some researchers log-transformed read counts in order to remove the discrete structure of the data and then applied several machine learning algorithms such as  $k$ -Nearest Neighbor ( $k$ NN), Support Vector Machines (SVMs) and logistic regression on log-transformed counts (Zararsiz et al., 2017a; Zararsiz et al., 2017b; Tan et al., 2014).

On the other hand, some researchers modelled the data more directly. For this purpose, discrete distributions two of which are Poisson and Negative Binomial (NB) are considered for RNA-Seq data modelling and data classification methods. Witten (2011) developed a classifier called Poisson Linear Discriminant Analysis (PLDA) assuming that RNA-Seq data comes from Poisson distribution. It is mentioned in Witten's paper that Poisson distribution based model can be extended to the Negative Binomial model in order to improve the classification performance when we have biological replicates in the data (Witten, 2011). The reason behind that is when there are biological replicates (multiple individuals) in the data; the variance of the data exceeds its mean. Therefore, the data is not distributed Poisson anymore. As a consequence, Dong et al. (2016) proposed a new classifier called Negative Binomial Linear Discriminant Analysis (NBLDA) for RNA-Seq data which assumes that data comes from Negative Binomial distribution.

When we have biological replicates in the data it is of importance to investigate the mean variance relationship of the counts before classification or linear modelling. Recently, mean variance modeling at the observational level (voom) method is proposed by Law et al. (2014) in order to estimate the mean variance

relationship of the log-transformed counts. It is shown that mean variance trend is more precise after voom transformation which enables Gaussian classification method become applicable (Zararsiz, 2015). Given these advantages, Zararsiz et al. (2017b) incorporated voom method into the Diagonal Linear Discriminant Analysis (DLDA) and proposed a new classifier called voomDLDA (voom based DLDA) for RNA-Seq data which may contain biological replicates.

Note here that all the aforementioned classification models assume that genes (measurement on the features) are independent (Dong et al. 2016; Witten, 2011; Zararsiz, 2015). However, since genes are highly correlated with each other on the same pathway or network, the strong independence assumption is not realistic for RNA-Seq problems and this may result in low performance in RNA-Seq data classification. Therefore, researchers have focused on the classification models for RNA-Seq data incorporating the dependency between genes. Sparse Quadratic Discriminant Analysis (SQDA) and Gaussian copula approach are recently proposed RNA-Seq classification methods of this type (Sun and Zhao, 2015; Zhang, 2017). Zhang (2017) modelled counts with multivariate Gaussian copula whereas Sun and Zhao (2015) modelled log-transformed counts with the multivariate normal distribution. It has shown that incorporating the dependence structure into the model increases the performance of RNA-Seq data classification.

In this thesis, we proposed a new classifier called quantile transformation Quadratic Discriminant Analysis (qtQDA) using RNA-Seq expression profiles. The proposed classifier incorporates dependence structure between genes into the classification model. There are three important contributions while classifying samples given their gene expression profiles. The first contribution is that instead of transforming the counts, e.g. using a log transformation, and modelling those log-transformed counts we transform the counts using a quantile transformation to be used in the classification of a new sample. The second contribution is to use a novel application of a powerful regularization technique for covariance matrix estimation. We use an R package called “corpcor” for this purpose (readers are referred to Shafer and Strimmer (2005) for more details). With the help of this package, we guarantee that estimated covariance matrix is

symmetric and positive definite and therefore can be used in the calculation of posterior probabilities. To the best of our knowledge, this is the first time this regularization technique is used in classification problems. The last contribution is the estimation of covariance matrix where we use two different techniques: simple covariance matrix estimation and local covariance matrix estimation. We claim that class-specific covariance matrices estimated by using local dependence function may improve the covariance matrix estimation and this may affect the classification performance of RNA-Seq data.

The rest of the thesis is organized as follows. In Chapter 2, some preliminaries are presented and some basic concepts are overviewed. In Chapter 3, some of powerful machine learning algorithms together with specialized RNA-Seq data classifiers are summarized. In Chapter 4, a new approach to RNA-Seq data classification, which is called qtQDA classifier, is given in details. Moreover, not only applications of the proposed method but also results are given within the same chapter. In chapter 5, a new local covariance matrix estimation technique is explained and then applied on real data sets. The results are analyzed under different covariance matrix estimations.

## CHAPTER 2: RNA SEQUENCING AND DISCRIMINANT ANALYSIS

In this chapter, we first introduce some basic concepts and important terms in molecular biology in order to understand RNA sequencing experiments. Then we explain RNA-Seq experiments and the structure of the data set obtained from the RNA sequencing experiments. After that, we give two discrete models, which are based on discrete distributions for RNA-Seq data sets. Lastly, we explain the discriminant analysis which can be used for any classification problems.

### *2.1 Basic Concepts and Terms*

**DNA** which stands for deoxyribonucleic acid is a molecule containing inheritable information of any organism. It stores all the genetic instructions which are necessary in the development, functioning and growth of all organisms. The genetic information is coded as a sequence of nucleotides: Adenine (A), Thymine (T), Cytosine (C) AND Guanine (G). DNA molecule consists of two strands that coil around each other in a double helix form (Watson and Crick, 1953). These two strands, each of which stores the same biological information are bonded together according to “base pairing rules” (A with T and C with G). See Figure 2.1

**RNA** which stands for ribonucleic acid is a single-stranded molecule and has important biological roles in the coding, decoding, regulation, carrying the information and expression of genes. It can be seen from the Figure 2.1 that it consists of the same nucleobases as DNA except Thymine which is displaced by Uracil (U). There are several types of RNA existing in the cell such as messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and micro RNA (miRNA) for different purposes.

**cDNA** is called complementary DNA which is synthesized from a single-stranded mRNA produced by reverse transcription.

**Gene** is a segment of DNA, which encodes a functional RNA or protein product, that is, it encodes the instructions for building proteins which performs the

biological functions of the cell (Slack, 2014).

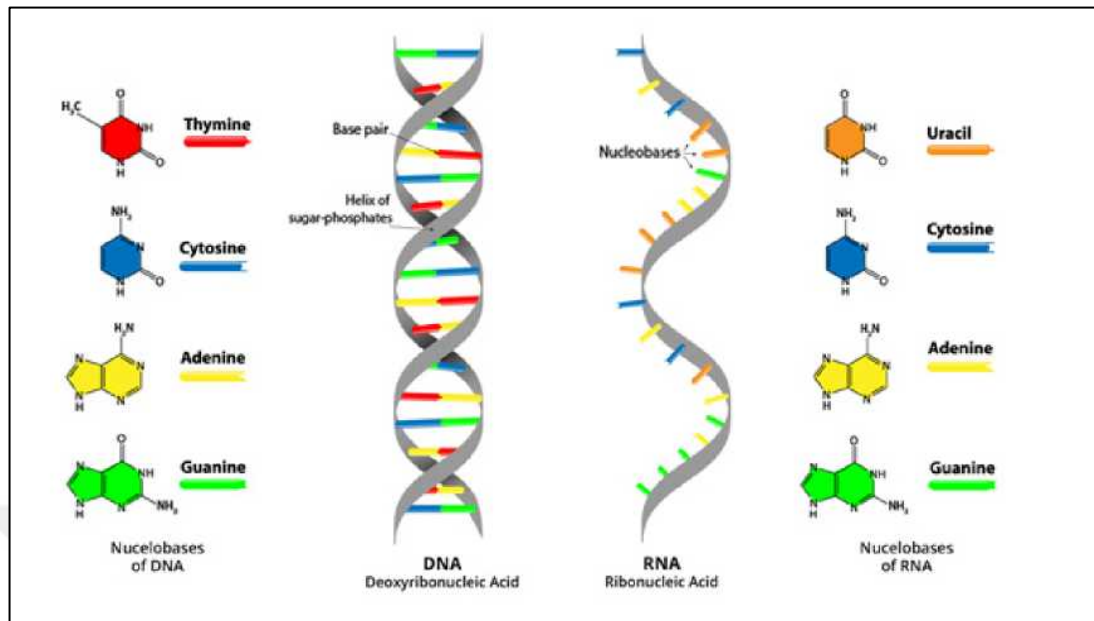


Figure 2.1. The detailed structure of DNA and RNA (Source: MacKenzie, 2010)

**Genome** is the genetic information contains all the hereditary information of an organism. The genome includes both coding and non-coding regions which are called exon and intron, respectively.

**Transcription** is the initial step of gene expression. Since DNA never leaves the nucleus, it creates an RNA in order to deliver the information stored in DNA to outside of the nucleus. Then a particular segment of DNA is copied into RNA by the RNA polymerase enzyme. The non-coding segments, known as introns, in RNA sequence are removed by splicing and the remaining coding sequences, known as exons, are combined together in the mRNA. See Figure 2.2.

**Translation** As soon as mRNA leaves the nucleus translation initiates with ribosome attachment to the mRNA molecule. Then mRNA is decoded by tRNA in order to produce a specific amino acid or polypeptide chain. See Figure 2.2

Technical replicates refer to repeated measurements that use the same biological sample (Figure 2.3). They are used to accurately measure the variability arises from the experiments that is, they tell us how accurately genes are measured. See

(Klaus, 2015) for more details.

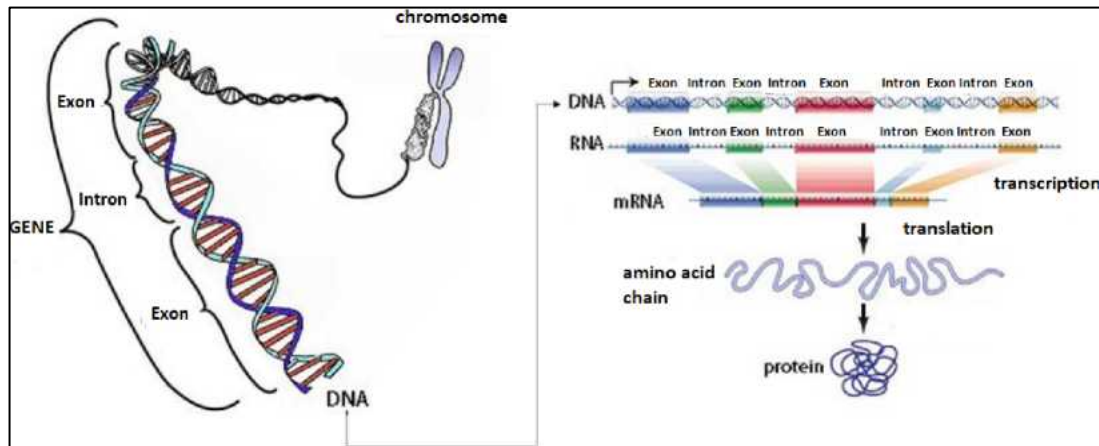


Figure 2.2. Transcription and translation processes (Source: Özdoğan, 2018)

**Biological replicates** refer to parallel measurements that use biologically different samples, which are used to measure the biological variation between different samples. See Figure 2.3. Using biological replicates in the experiments enables us to determine whether the experimental effect is biologically relevant or not. See (Klaus, 2015) for more details.

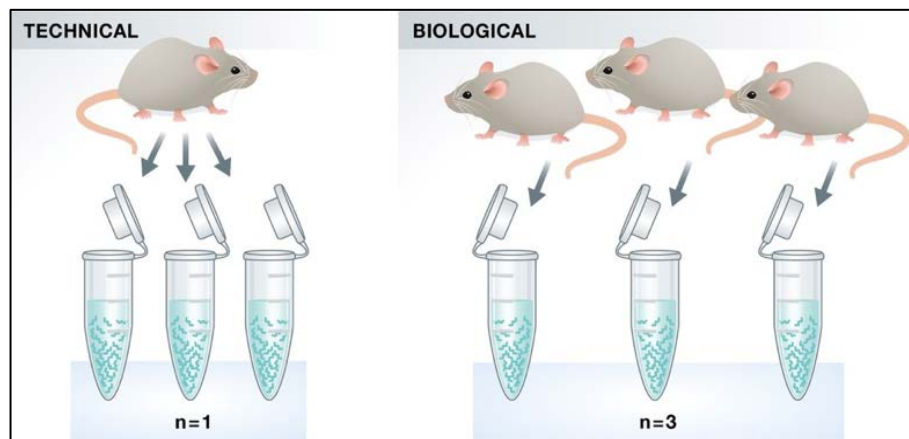


Figure 2.3. Technical replicates vs biological replicates (Source: Klaus, 2015)

## 2.2 Next Generation RNA Sequencing

It is known that there exists nearly 25,000 genes in the human genome. The expression level of all these genes can be measured simultaneously using a variety of sophisticated techniques. Currently the most popular technique is next-

generation RNA sequencing which was first introduced in 2008 (Mortazavi et al., 2008; Holt and Jones, 2008). Since then the popularity of RNA-Seq has increased due to the many advantages such as unprecedented sequencing speed, cost-effectiveness, accuracy in genomic analysis, high resolution etc. (AbuElQumsan, 2018; Mardis 2008; Wang et al., 2009; Metzker, 2010).

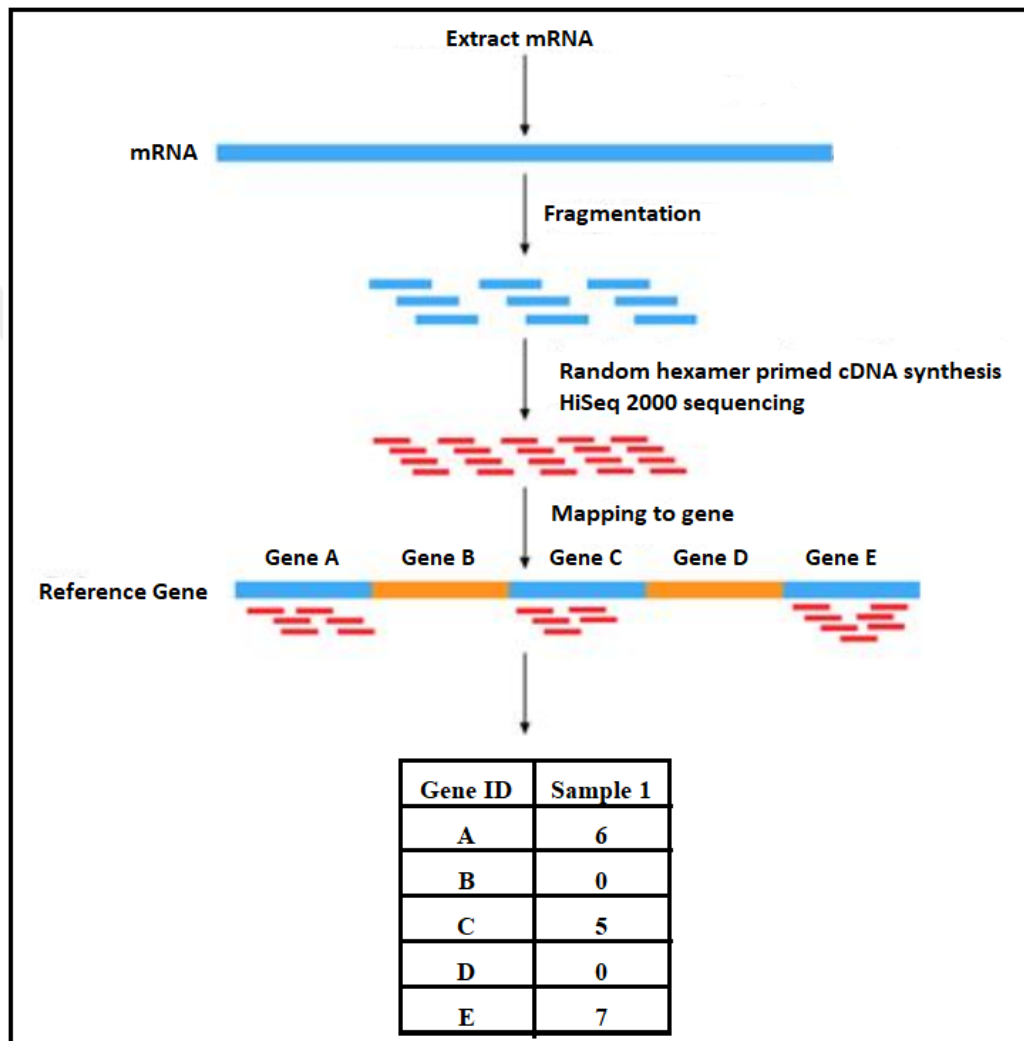


Figure 2.4. A workflow of RNA Sequencing (Source: adopted from University of Gothenburg Institute of Biomedicine Lecture Notes)

The first step in the technique is to convert all RNA samples to be sequenced into cDNA fragments (a cDNA library). Adapters are then attached to the end of the fragments which include functional elements that allow sequencing. The cDNA library is then analyzed by NGS and short sequences (reads) are produced. Sequencing depth (the depth to which the library is sequenced) varies

depending on techniques which the output data will be used for. Then those reads are aligned and mapped to a reference gene/genome. After all, the number of reads mapped to corresponding gene is recorded as an output in the data matrix. See Figure 2.4 for the RNA-Seq workflow.

### 2.3 RNA-Seq Data

After RNA sequencing process, we obtain  $p \times n$  dimensional count matrix  $X = \{x_{ij} | 1 \leq i \leq p, 1 \leq j \leq n\}$  given in Figure 2.5. Note here that each row indicates the expression patterns of genes and each column indicates the expression profiles of samples, and  $x_{ij}$  is the expression level of  $i^{th}$  gene in the  $j^{th}$  sample. It can be seen from Figure 2.5 that each gene is denoted by a row vector, i.e.,  $X_i = [x_{i,1} \ x_{i,2} \ \dots \ x_{i,n}]$  and each sample is denoted by a column vector, i.e.,  $X_j = [x_{1,j} \ x_{2,j} \ \dots \ x_{p,j}]^T$ , where  $T$  denotes the transpose of a matrix. The total number of counts for sample  $j$ , which is defined as the library size is denoted by  $X_{\cdot j}$  where  $X_{\cdot j} = \sum_{i=1}^p x_{ij}$ . The total number of counts mapped to  $i^{th}$  gene is represented by  $X_{i\cdot}$  where  $X_{i\cdot} = \sum_{j=1}^n x_{ij}$  and the total sum of the number of counts within the data is denoted by  $X_{\cdot\cdot}$  where  $X_{\cdot\cdot} = \sum_{i=1}^p \sum_{j=1}^n x_{ij}$ . Moreover, there exists another term called size factor for the  $j^{th}$  sample,  $s_j$  and it can be estimated by  $s_j = \frac{X_{\cdot j}}{X_{\cdot\cdot}}$  in order to scale read counts due to the different sequencing depth.

	Sample 1	Sample 2	...	Sample $j$	...	Sample $n$	
$X =$	Gene 1	$x_{11}$	$x_{12}$	...	$x_{1j}$	...	$x_{1n}$
		$x_{21}$	$x_{22}$	...	$x_{2j}$	...	$x_{2n}$
	:	:	:	\ddots	:	\ddots	:
	Gene $i$	$x_{i1}$	$x_{ip}$	...	$x_{ij}$	...	$x_{in}$
	:	:	:	\ddots	:	\ddots	:
	Gene $p$	$x_{p1}$	$x_{p2}$	...	$x_{pj}$	...	$x_{pn}$

Figure 2.5. Gene Expression Data Matrix

Let us now consider the cervical cancer data (Witten, 2010) represented in Table



2.1 as an example for an RNA-Seq data matrix. The RNA-Seq data matrix of cervical cancer which is composed of read counts is obtained after some pre-processing. It can be seen from the Table 2.1 that there are 58 samples. The first 29 of the samples are tumor while the rest of the samples are non-tumor and each sample consists of 714 genes. The first element of the matrix,  $x_{11} = 865$  is the number of reads of gene “let-7a” for the first sample. Moreover, this first sample is non-tumor and  $X_{\cdot 1} = 22,449$  represents the library size of the first sample (Sample 1), which is the total number of reads for that sample.

Table 2.1. RNA-Seq data matrix of cervical cancer data (Source: Zararsiz, 2015).

miRNA	Non-tumor samples					Tumor samples					Total
	Sample-1	Sample-2	Sample-3	...	Sample-29	Sample-30	Sample-31	Sample-32	...	Sample-58	
let-7a	865	810	5,505	...	38	3,343	4,990	5,193	...	1,422	284,257
let-7g	447	173	1,922	...	126	737	4,141	3,760	...	2,081	128,348
miR-125b	1,038	5,007	2,595	...	106	381	1,463	201	...	1	337,394
miR-18a	5	4	10	...	0	10	9	56	...	0	634
miR-29a	320	447	904	...	4	413	4,619	1,398	...	1	134,382
miR-490-5p	0	1	0	...	0	0	0	0	...	0	19
miR-874	2	4	9	...	0	4	0	3	...	0	509
...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...
miR-93	10	18	126	...	0	36	133	211	...	0	7,028
miR-99b	24	92	97	...	177	36	62	19	...	20	11,718
<b>Total</b>	22,449	39,798	71,717	...	8,034	58,362	431,247	84,850	...	16,338	13,701,148

## 2.4 Notations

In this section we introduce some notations which will be utilized throughout the thesis.

- $\mathbf{X}$ :  $p \times n$  dimensional data matrix where  $p$  and  $n$  are the number of genes and samples, respectively.
- $X_{ij}$ : number of reads/counts mapped to gene  $i$  in sample  $j$ .
- $\mathbf{X}^{(k)} = [X_1^{(k)} \quad X_2^{(k)} \quad \dots \quad X_p^{(k)}]^T$ : a data matrix from the  $k$ th class where  $X_i^{(k)}$  is the vector of counts for the  $i$ th gene from the  $k$ th class.
- $X_{ij}^{(k)}$ : the number of reads that mapped to  $i$ th gene in the  $j$ th sample belongs to the  $k$ th class.

- $\mathbf{X}_j = [X_{1j} \ X_{2j} \ \dots \ X_{pj}]$ : the vector of counts of each gene for the  $j$ th sample or the transpose of the  $j$ th column of the data matrix  $\mathbf{X}$ .
- $K$ : the number of distinct classes (biological conditions).
- $\mathbf{Y} = \{Y_j: j = 1, \dots, n\}$  is an  $n$ -dimensional vector containing class labels of  $n$  observations.
- $Y_j \in \{1, \dots, K\}$  the class label of  $j$ th sample.
- $C_k = \{j: Y_j = k\}$ : the index set of samples from the  $k$ th class.
- $\mathbf{x}^*$ : a new observation to be classified.
- $Y^*$ : unknown class label of  $\mathbf{x}^*$  that will be predicted.
- $\pi_k$  the prior probability of an observation belonging to the class  $k$ .
- $\Sigma^k$ : variance-covariance matrix for the  $k$ th class.
- $L_j = \sum_i X_{ij}$ : library size for the  $j$ th sample or total counts across all genes in the  $j$ th sample.

## 2.5 Discrete Models Proposed for RNA-Seq Data

Due to the discrete structure of RNA-Seq data, researchers have considered modelling RNA-Seq data under the assumption that counts follow a family of discrete distributions. In this section of thesis, we give two models where counts are assumed to be drawn from Poisson and Negative Binomial distributions accordingly.

### 2.5.1 Poisson Model

Poisson based model was the first discrete model considered for RNA-Seq studies (Marioni et al., 2008, Bullard et al., 2010, Wang et al., 2010; Witten et al., 2010). Since biological replicates may cause overdispersion in the data, Poisson based models are only valid when there is only technical replicates (no biological replicates) in the data. Poisson based models assume that counts are drawn from Poisson distribution:

$$X_{ij} \sim \text{Poisson}(\mu_{ij}), \mu_{ij} = s_j g_i \quad (2.1)$$

where  $g_i$  is total number of counts for  $i$ th gene and  $s_j$  is the total number of counts for  $j$ th sample satisfying  $\sum_{j=1}^n s_j = 1$ . It is obvious that

$$E(X_{ij}) = \text{Var}(x_{ij}) = \mu_{ij}$$

### 2.5.2 Negative binomial model

Poisson distribution can be implemented when there are only technical replicates in the data. However, when we have biological replicates in the data variance exceeds mean that is data becomes overdispersed. Therefore, another model, a negative binomial model, has been proposed (Robinson, McCarthy and Smyth, 2010; Anders and Huber, 2010). Moreover, it has been proved in (Dong et al., 2016) that if there are no biological replicates then the data follows Poisson distribution. Therefore, we assume that marginal distributions of the counts are negative binomial which is given as:

$$X_{ij} \sim \text{NB}(\mu_{ij}, \Phi_i), \mu_{ij} = s_j g_i \quad (2.2)$$

where  $\mu_{ij}$  and  $\Phi_i$  are the mean parameter of  $i$ th gene in the  $j$ th sample and dispersion parameter for gene  $i$ , respectively. Then, expectation and variance can be easily calculated as:

$$E(X_{ij}) = \mu_{ij}$$

$$\text{Var}(x_{ij}) = \mu_{ij} + \mu_{ij}^2 \Phi_i > \mu_{ij}$$

It is obvious to see that variance of the counts exceeds mean when we assume that the counts are marginally negative binomial which is the case when we have biological replicates in the data. Now assume that there are no biological replicates in the data. Then the dispersion will be zero and the mean of the counts will be equal to variance of the counts, i.e counts will be marginally Poisson distributed.

## 2.6 Discriminant Analysis

Discriminant analysis is a statistical technique that is used to analyze data when we have categorical dependent variable such as types of a disease (e.g. types of breast cancer). Linear Discriminant Analysis (LDA) is the generalization of Fisher's linear discriminant analysis which was originally proposed by Fisher (1936). There exist many applications of LDA in various fields such statistical data analysis, pattern recognition and machine learning (Duda et al., 2001). It can be used either for dimensionality reduction or classification. For the classification purpose, the objective is to develop a discriminant function that can differentiate between classes (dependent categorical variable). Thus, it enables us to observe significant differences among the classes.

Let  $X$  be a random variable for the data and  $Y$  be a random variable for the class labels. Assume that we have  $k \in \{1, 2, \dots, K\}$  different classes with prior probabilities  $\pi_k$  such that  $\sum_{k=1}^K \pi_k = 1$ . Assume that  $f_k(x)$  is class conditional density of a sample  $x$  that belongs to the class  $k$  and  $n_k$  is the number of observations in the  $k$ th class. In order to assign a new observation,  $\mathbf{x}^*$  to one of  $K$  distinct classes, posterior probabilities  $\Pr(Y = k|X = \mathbf{x})$  are required to be computed or estimated. Hence, by Bayes theorem, the posterior probabilities can be estimated as follows:

$$\Pr(Y = k|X = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{j=1}^K f_j(\mathbf{x})\pi_j} \quad (2.3)$$

which is equivalent to say that

$$\Pr(Y = k|X = \mathbf{x}) \propto f_k(\mathbf{x})\pi_k \quad (2.4)$$

The equation (2.4) is called Bayes' rule.

According to Linear Discriminant Analysis (LDA), it is assumed that class conditional density is multivariate normal, i.e.

$$f_k(\mathbf{x}) = \Pr(Y = k|X = \mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma^{-1}(\mathbf{x} - \mu_k)\right\} \quad (2.5)$$

where  $\mu_k$  represents the class-specific mean vector and  $\Sigma$  represents covariance matrix. After replacing Equation (2.5) into the Equation (2.3) and applying some algebra, we get the following linear discriminant function:

$$\delta_k^{\text{LDA}}(\mathbf{x}^*) = (\mathbf{x}^*)^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k \quad (2.6)$$

where

$$\hat{\mu}_k = \sum_{j=1}^{n_k} \frac{\mathbf{x}_j}{n_k} \quad (\text{sample mean vector for class } k)$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^K \sum_{j=1}^{n_k} (\mathbf{x}_j - \hat{\mu}_k)(\mathbf{x}_j - \hat{\mu}_k)^T \quad (\text{sample pooled variance-covariance matrix})$$

$\hat{\pi}_k = \frac{n_k}{n}$  where  $n_k$  is the number of observations in class  $k$  and  $n$  is the total number of samples in the training.

The parameters are estimated from the training data set by applying Maximum Likelihood Estimator. Then, a new observation  $\mathbf{x}^*$  is assigned to the class which maximizes the discriminant function given in Equation (2.6).

Note here that if we assume genes are independent which means covariances or off diagonals are assumed to be zero then we only need to estimate the diagonal covariance matrices. Then the resulting discriminant function is called Diagonal Linear Discriminant Analysis (DLDA) and given by

$$\delta_k^{\text{DLDA}}(\mathbf{x}^*) = -\sum_{i=1}^p \frac{(\mathbf{x}_i^* - \bar{x}_{ik})^2}{s_i^2} + 2 \log \hat{\pi}_k \quad (2.7)$$

where  $\mathbf{x}_i^*$  is the  $i$ th element of the new observation  $\mathbf{x}^*$ ,  $s_i^2$  is sample variance of  $i$ th gene and  $\bar{x}_{ik}$  is sample mean of  $i$ th gene in  $k$ th class.

Quadratic Discriminant Analysis (QDA) is the generalization of LDA where we

assume that covariance matrices are different for each class. Therefore, it is essential to estimate covariance matrices separately for each class. Since we estimate class-specific covariance matrices and we incorporate the dependence structure into the model using class-specific covariance matrices, we use QDA in the proposed model. This is important because class-specific covariance matrices together with the dependence structure may contain particular information for each class which may increase the performance of the classification.



## CHAPTER 3: DISCRETE AND MACHINE LEARNING METHODS FOR RNA-SEQ DATA CLASSIFICATION

In this chapter, we summarize not only a number of machine learning algorithms, which are widely applied on RNA-Seq data but also discrete model based statistical algorithms, which are only proposed for RNA-Seq data classification.

### *3.1 Machine Learning Methods Applied on RNA-Seq Data*

#### *3.1.1 k-Nearest Neighbour Algorithm*

$k$ -Nearest Neighbor ( $k$ NN) algorithm is one of the mostly applied supervised machine learning algorithms used in variety of applications such as face recognition, pattern recognition, bioinformatics, etc. (Altman, 1992; Yao and Ruzzo, 2006). There are many advantages of  $k$ NN. For instance, it is one of the simplest and easiest machine learning algorithms that can be implemented on many data sets. However, there exist many disadvantages of  $k$ NN such as high memory requirement (since it stores almost all of the training data) and being sensitive to outliers. Additionally, it can perform pretty slowly when the dimension of the data increases.

Basically,  $k$ NN algorithm is based on distance which can be also seen as similarity or proximity in the literature. Thus, it is crucial to find out the appropriate distance to be used in the algorithm. Although Euclidean distance is the most commonly used distance in  $k$ NN there are other distances such as Hamming, Pearson, Mahalanobis distances that can be used in  $k$ NN algorithm depending on the structure of the data.

$k$  in  $k$ NN represents the number of points closest to the new point that has to be classified. For instance, if  $k = 3$  then the algorithm selects three closest points in the neighborhood of the new point (green dot in Figure 3.1) and assigns the new point to the class of closest point (red triangles in Figure 3.1).

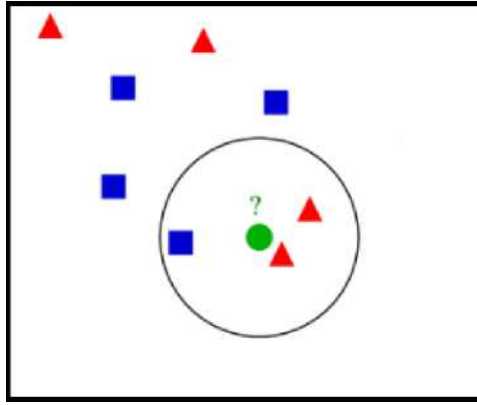


Figure 3.1. An example of  $k$ NN algorithm when  $k = 3$

### 3.1.2 Support Vector Machines

Support Vector Machine (SVM), which was developed by Vapnik (2000) is not only a powerful but also a widely used supervised machine learning algorithm. It has been applied to quite a few areas such as pattern recognition, image processing, text categorization, medicine, biological sciences, etc.

What SVM does is to identify an ideal decision boundary in order to separate classes and this ideal decision boundary, which is also called ideal separation hyper-plane is determined according to the maximum margin principle. In other words, the algorithm chooses the decision boundary which maximizes the distance between classes according to the maximum margin principle. The vectors which define the hyper-planes are called support vectors.

If the data set we are working on is linearly separable then SVM performs very efficiently and produces a hyper-plane that utterly splits the vectors into two classes, which do not overlap. However, perfect separation is not always the case which may result in producing many possible hyper-planes. In that situation, SVM searches for the hyper-plane, which minimizes misclassification rate and maximizes the margin simultaneously (Sayad, 2019). If the data set is linearly non-separable then SVM utilizes kernel functions which transform the data into a high dimensional feature space to make it linearly separable. See Figure 3.2.



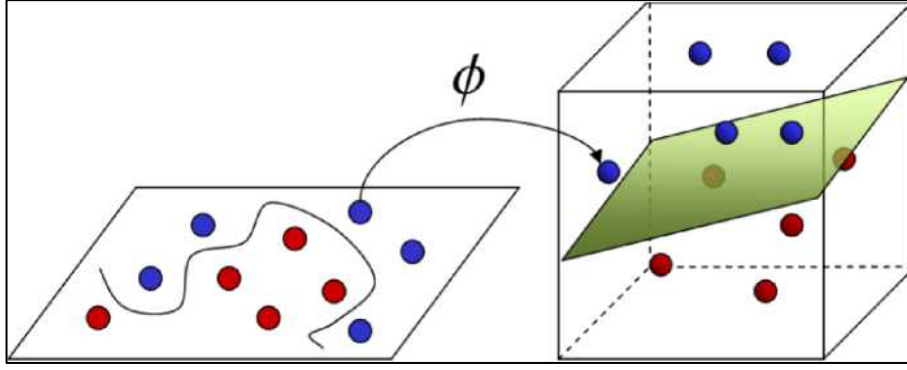


Figure 3.2. Kernel method (Source: Wilimitis, 2018)

Assuming that  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are two vectors then two commonly applied kernel types, polynomial and Gaussian radial basis kernels, are defined as follows:

- Polynomial:  $\phi(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$  where  $d$  is the degree of the polynomial
- Gaussian radial basis:  $\phi(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$  where  $\gamma > 0$ .

### 3.1.3 Logistic Regression Classifier

Logistic regression simply is a statistical model which uses a logistic function to propose a prediction model for a binary dependent variable with two possible values such as “1” and “2”. We note here that it can be further extended to the cases where the dependent variable has more than two possible values. Logistic regression has many applications in many areas such as machine learning, engineering, economics, medicine, etc. We mentioned earlier that we will focus on binary classification problems. Thus, we explain binary logistic regression which uses binomial distribution. For more details we refer to (Tan, 2014; Friedman, 2010).

Given the notations in Section 2.4, assume that  $K = 2$ . Then class-conditional probabilities that  $j$ th observation  $\mathbf{X}_j$  belongs to class “1” and “2” are modeled as

$$\Pr(Y_j = 1 | \mathbf{X}_j) = \frac{1}{1 + e^{\alpha + \mathbf{x}_j^T \beta}}; \Pr(Y_j = 2 | \mathbf{X}_j) = \frac{e^{\alpha + \mathbf{x}_j^T \beta}}{1 + e^{\alpha + \mathbf{x}_j^T \beta}} \quad (3.1)$$

Where  $\alpha$  and  $\beta$  are unknown scalar and vector of length  $p$  respectively. Equation (3.1) can also be written as “logistic” or “log-odds” transformation:

$$\log \frac{\Pr(Y_j=1|X_j)}{\Pr(Y_j=2|X_j)} = \alpha + \mathbf{X}_j^T \beta \quad (3.2)$$

Then the corresponding log-likelihood is given as follows:

$$l(\alpha, \beta) = \sum_{j=1}^n \left\{ c_j (\alpha + \mathbf{X}_j^T \beta) - \log \left( 1 + e^{\alpha + \mathbf{X}_j^T \beta} \right) \right\} \quad (3.3)$$

where  $c_j = I\{Y_j = 1\}$ . One can apply MLE in order to estimate the unknown parameters  $\alpha$  and  $\beta$ . The maximum likelihood estimates of  $\alpha$  and  $\beta$ ,  $\hat{\alpha}$  and  $\hat{\beta}$ , are obtained by maximizing the Equation (3.3). Once unknown parameters are estimated a new observation is then assigned to a particular class according to the following probability function:

$$\Pr(Y^* = 1|\mathbf{x}^*) = \frac{1}{1 + e^{\hat{\alpha} + \hat{\beta}^T \mathbf{x}^*}} \quad (3.4)$$

Generally, a cutoff point,  $0 \leq t \leq 1$ , is chosen in order to assign the new observation into one of the classes, roughly speaking we assign the new observation into the class 1 if  $\Pr(Y^* = 1|\mathbf{x}^*) > t$  and into the class 2 otherwise.

Indeed, logistic regression classifier is not applicable in high dimensional data classification problems, as the parameter estimation becomes unstable and may contain infinite elements (Tan et al., 2014). Hence, to overcome this problem some regularized version of logistic regression classifiers are developed. One way is to regularize Equation (3.3) by adding a penalty term, which is a convex penalty function  $P(\beta)$ . The penalized log-likelihood function is defined as follows:

$$l(\alpha, \beta) = \sum_{j=1}^n \left\{ c_j (\alpha + \mathbf{X}_j^T \beta) - \log \left( 1 + e^{\alpha + \mathbf{X}_j^T \beta} \right) \right\} - \lambda P(\beta) \quad (3.5)$$

where  $\lambda$  is a non-negative tuning parameter and can be chosen by cross validation.

We note here that ridge and lasso penalties are two commonly applied penalties in the literature.  $P(\beta) = \|\beta\|^2$  is a ridge penalty (Hoerl and Kennard, 1970) and  $P(\beta) = \|\beta\|_1$  is a lasso penalty (Tibshirani, 1996). We also note that GLMnet package uses the lasso penalty in the log-likelihood function.

### 3.2 Discrete Methods Applied on RNA-Seq Data

#### 3.2.1 Poisson Linear Discriminant Analysis

Poisson Linear Discriminant Analysis (PLDA) was proposed by Witten (2011) for RNA-Seq data classification. The model assumes that genes are independent and counts are marginally Poisson distributed as described before in Section 2.5.1.

Since samples are drawn from  $K$  distinct classes then (2.1) is extended where the class specific Poisson model for RNA-Seq data is defined as follows:

$$(X_{ij} | Y_j = k) \sim \text{Poisson}(\mu_{ij}d_{ki}), \mu_{ij} = s_j g_i \quad (3.6)$$

where  $s_j$  is the size factor for the  $j$ th sample,  $g_i = X_i$  is the total number of read counts for the  $i$ th gene and  $d_{ki}$  is the term permits  $i$ th gene to be differentially expressed in the  $k$ th class. Then the probability mass function of  $X_{ij} = x_{ij}$  in model (3.6) is

$$\Pr(X_{ij} = x_{ij} | Y_j = k) = e^{-\mu_{ij}d_{ki}} \frac{(\mu_{ij}d_{ki})^{x_{ij}}}{x_{ij}!} \quad (3.7)$$

Since genes are assumed to be independent in PLDA, by substituting (3.7) in Bayes' rule given in (2.4) one can obtain the following discriminant function for PLDA:

$$\begin{aligned}
\Pr(Y^* = k | \mathbf{x}^*) &= \log \left( e^{-s^* g_i d_{ki}} \frac{(s^* g_i d_{ki})^{x_i^*}}{x_i^*!} \pi_k \right) \\
&= \sum_{i=1}^p -s^* g_i d_{ki} + \sum_{i=1}^p x_i^* \log(s^* g_i d_{ki}) + \log \pi_k - \sum_{i=1}^p \log x_i^*
\end{aligned}$$

which can be simplified to the following discriminant function:

$$\delta_k^{\text{PLDA}}(\mathbf{x}^*) = \sum_{i=1}^p x_i^* \log d_{ki} - s^* \sum_{i=1}^p g_i d_{ki} + \log \pi_k \quad (3.8)$$

where  $s^*$  is the size factor for the new observation. A new observation  $\mathbf{x}^*$  is then assigned to the class which maximizes the discriminant score given in (3.8). See (Witten, 2011) for details and for the estimation of the parameters included in the PLDA discriminant function.

### 3.2.2 Negative-Binomial Linear Discriminant Analysis

It is known that RNA-Seq data follows negative binomial distribution when there exist biological replicates in the data set (Witten, 2011; Dong et al., 2016; Zararsiz et al., 2017a; Robinson and Smyth, 2008). The reason for this assumption is the fact that the variability of biological replicates leads to overdispersion in the data, which means variance of the data exceeds mean of the data in most of the cases. Given this information, Dong et al. (2016) introduced a new linear discriminant analysis, Negative Binomial Linear Discriminant Analysis (NBLDA), which uses the negative binomial distribution (see Section 2.5.1) as class conditional densities instead of Poisson distribution to cope with the overdispersion arising from the biological replicates.

Since samples are drawn from  $K$  different classes, (2.2) is extended where the class specific negative binomial model for RNA-Seq data is defined as follows:

$$(X_{ij} | Y_j = k) \sim \text{NB}(\mu_{ij} d_{ki}, \Phi_i), \mu_{ij} = s_j g_i \quad (3.9)$$

where  $s_j$  is the size factor for the  $j$ th sample,  $d_{ki}$  is the term permits  $i$ th gene to

be differentially expressed in the  $k$ th class,  $g_i = X_i$ . and  $\Phi_i > 0$  is the total number of reads and dispersion parameter for the  $i$ th gene, respectively. Then the probability mass function of  $X_{ij} = x_{ij}$  in model (3.9) is

$$\Pr(X_{ij} = x_{ij} = k | Y_j = k) = \frac{\Gamma(x_{ij} + \Phi_i^{-1})}{x_{ij}! \Gamma(\Phi_i^{-1})} \left( \frac{s_j g_i d_{ki} \Phi_i}{1 + s_j g_i d_{ki} \Phi_i} \right)^{x_{ij}} \left( \frac{1}{1 + s_j g_i d_{ki} \Phi_i} \right)^{\Phi_i^{-1}} \quad (3.10)$$

Since genes are assumed to be independent in PLDA, by replacing (3.10) into (2.4) (Bayes' rule) we get the following discriminant function for NBLDA:

$$\begin{aligned} \delta_k^{\text{NBLDA}}(\mathbf{x}^*) &= \sum_{i=1}^p x_i^* [\log d_{ki} - \log(1 + s^* g_i d_{ki} \Phi_i)] \\ &\quad - \sum_{i=1}^p \Phi_i^{-1} \log(1 + s^* g_i d_{ki} \Phi_i) + \log \pi_k \end{aligned} \quad (3.11)$$

where  $s^*$  is the size factor for the new observation. The new observation,  $\mathbf{x}^*$  is then assigned to the class that maximizes the discriminant function, Equation (3.11) The dispersion parameter in NBLDA classifier is estimated by a shrinkage method which is proposed by Yu et al. (2013). See (Yu et al., 2013; Dong et al., 2016) for more details about NBLDA and the estimation of dispersion parameter.

### 3.2.3 Voom Based Diagonal Linear Discriminant Analysis

Voom is a variance modelling at the observational level method which was proposed by Law et al. (2014). Voom estimates the mean variance relationship from the log transformed counts. Additionally it provides precision weights (Ritchie et al., 2015) for the downstream analysis. It is shown that voom method has a number of advantages such as performing the best in controlling the type-I error and giving lowest false discovery rate. In order to take the advantage of the voom method in RNA sequencing data classification, Zararsiz et al. (2017b) integrated the voom method into DLDA classifier given in Equation (2.7) and they called the new classifier as voomDLDA.

Differently from PLDA and NBLDA, voomDLDA has the following discriminant score (Zararsiz et al., 2017b):

$$\delta_k^{\text{voomDLDA}}(\mathbf{x}^*) = -\sum_{i=1}^p \frac{(z_i^* - \bar{z}_{wik})^2}{s_{wi}^2} + 2 \log \pi_k \quad (3.12)$$

The notations used in (3.12) are as follows:

$z_i^*$ : log-cpm values for the new observation defined by

$$z_i^* = \log_2 \left( \frac{x_i^* + 0.5}{X_{.*} + 1} \times 10^6 \right)$$

where  $X_{.*}$  is the library size for the new observation

$s_{wi}^2$ : weighted pooled variance of  $i$ th gene where  $n_k$  is the number of observations in  $k$ th class and it is calculated as follows:

$$s_{wi}^2 = \sum_{j=1}^K (n_k - 1) s_{wik}^2 / (n - k)$$

$s_{wik}^2$ : weighted variance of  $i$ th gene in  $k$ th class where  $j_k, \dots, j_{k+1} - 1$  belong to the  $k$ th class and it is evaluated as follows:

$$s_{wik}^2 = \frac{\sum_{j=j_k}^{j_{k+1}-1} w_{ij}}{\left( \sum_{j=j_k}^{j_{k+1}-1} w_{ij} \right)^2 - \sum_{j=j_k}^{j_{k+1}-1} w_{ij}^2} \sum_{j=j_k}^{j_{k+1}-1} w_{ij} (z_{ij} - \bar{z}_{wik})^2$$

where  $w_{ij}$  are estimated precision weights

$\bar{z}_{wik}$ : the class specific weighted mean for the class  $k$ , which is defined by

$$\bar{z}_{wik} = \frac{\left( \sum_{j=j_k}^{j_{k+1}-1} w_{ij} z_{ik} \right)}{\sum_{j=j_k}^{j_{k+1}-1} w_{ij}}$$

## CHAPTER 4: THE MODEL

In this chapter, we explain the proposed model, qtQDA, in details. We also give the implementation of the proposed model on real RNA-Seq data sets and compare the proposed model with a number of existing classifiers in the literature.

### 4.1 General Structure of The Model

qtQDA is a new classifier for RNA-Seq data based on a model where genes are assumed to be dependent and marginally negative binomial distributed as it is given in Section 2.5.1. The proposed classifier integrates quantile transformation with QDA in order to incorporate the dependency between genes into the model.

The main steps of the qtQDA classification algorithm are given as follows:

- i. **Preprocess:** Filter the RNA-Seq data and get the data ready for the next steps of the algorithm.
- ii. **Gene Selection:** Select most informative genes to be used for downstream analysis using edgeR pipeline as explained in Section 4.2.2.
- iii. **Parameter Estimation I:** Estimate the parameters of negative binomial marginal of each gene selected in the second step of the algorithm using “estimateDisp” function from *edgeR* package. The parameters are estimated using Generalized Linear Model(s) (GLMs) which is described in Section 4.2.3.
- iv. **Quantile transformation:** Use inverse quantile transformation as elaborated in the Section 4.2.4 to transform negative binomially distributed variable to the multivariate normally distributed variable with mean  $\mathbf{0}$  and class-specific variance-covariance matrices  $\Sigma^k$ .
- v. **Parameter Estimation II:** Estimate the class-specific covariance

matrices and apply a regularization technique to guarantee that covariance matrices are symmetric and positive definite See Section 4.2.5 for more details.

vi. **Classification step:**

- ✓ Apply inverse quantile transformation to a new observation given in Section 4.2.6 and develop a discriminant function based on QDA.
- ✓ Assign the new observation to the class that gives the highest discriminant score which is developed in Section 4.2.6.

These steps are repeated until all new observations are classified.

## 4.2 *qtQDA model*

We assume that counts follow negative binomial distribution which is given in Section 2.5.1. Then the class specific model for negative binomially distributed counts is defined as follows:

$$X_{ij}^{(k)} \sim \text{NB}(\mu_{ij}^{(k)}, \Phi_i^{(k)}) \quad (4.1)$$

where  $\mu_{ij}^{(k)}$  and  $\Phi_i^{(k)}$  are class specific mean for  $i$ th gene in the  $j$ th sample and class specific dispersion for  $i$ th gene.

We now explain the steps of the algorithm in the following sections.

### 4.2.1 Preprocess

Since genes with low counts can effect the classification algorithm, genes with very low counts across all libraries have been filtered. Only genes that are expressed at a count per million (cpm) above 0.025 (median library size in cpm) in at least 453 samples (average counts for each gene across all samples with library size above 0.025 cpm) are considered in the downstream analysis. After filtering, number of genes reduced from 52580 to 16296 for HapMap data and



62706 to 36841 for Prostate cancer data. Since cervical cancer data only has 714 genes, we kept all genes for the downstream analysis.

#### **4.2.2 Gene Selection**

It is mentioned previously that tens of thousands of genes are measured simultaneously with the advent of RNA-Seq technology. However, it is essential to select the significant genes, which are informative for the purpose of class prediction. Although Witten (2011) introduced a screening method in order to detect informative genes it is only applicable to counts that has Poisson marginals. Since we assume that the counts are marginally negative binomial, we apply edgeR pipeline (Robinson et al., 2008; Robinson et al., 2010) which is suitable for negative binomially distributed RNA-Seq data to select distinguishing genes. What edgeR does is to pursue the following process: (1) filter genes with very low counts across samples; (2) perform a Likelihood Ratio Test (LRT) to the remaining genes in order to test for Differentially Expressed (DE) genes; (3) make a list of DE genes where DE genes are sorted by LRT score; (4) select the top  $p$  genes from the list.

#### **4.2.3 Parameter Estimation I**

In order to apply the proposed classifier in practice we first need to estimate the parameters of negative binomial marginals (i.e. means and dispersions) for all genes selected in the previous step of the algorithm and for each class, accordingly.

Means of genes are estimated by the methodology given in edgeR package which uses Generalized Linear Models (GLMs). GLMs are extension of classical linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). The reason for applying GLM theory is the fact that it can correctly specify mean-variance relationship for read counts (Mccarthy et al., 2012). Note to mention that it is considerably fast.

Let  $X_{ij}^{(k)} \sim \text{NB}(\mu_{ij}^{(k)}, \Phi_i^{(k)})$  denotes the number of reads mapped to gene  $i$  in the  $j$ th sample from the  $k$ th class. Then for each gene, GLM theory is used in order to fit a generalized log-linear model:

$$\log \mu_{ij}^{(k)} = D_j^T \alpha + \log L_j = [0 \quad \dots \quad 1 \quad 0 \quad \dots \quad 0] \begin{bmatrix} \alpha_i^{(1)} \\ \vdots \\ \alpha_i^{(k)} \\ \vdots \end{bmatrix} + \log L_j = \alpha_i^{(k)} + \log L_j \quad (4.2)$$

and  $D_j^T$  is the  $j$ th row of the design matrix which can also be seen vector of covariates (groups or classes),  $\alpha$  is a vector of coefficients coming from regression and  $L_j = X_{.j}$  is library size for the  $j$ th sample (the total read counts across all genes in the sample). The estimate for means are then obtained by

$$\hat{\mu}_{ij}^{(k)} = L_j \exp(\hat{\alpha}_i^{(k)}).$$

Dispersion parameter for each gene is estimated by using the Cox-Reid Adjusted Profile Likelihood (APL) function (Cox and Reid, 1987; McCarthy et al., 2012; Chen et al., 2014) which is given as follows:

$$\text{APL}_i(\Phi_i^{(k)}) = l(\Phi_i^{(k)}) - \frac{1}{2} \log \det(I_i^{(k)}) \quad (4.3)$$

where  $l(\cdot)$  represents log-likelihood function and  $I_i^{(k)}$  represents the Fisher information matrix of  $\hat{\alpha}_i^{(k)}$ . One can simply maximize the Equation (4.3) in order to estimate dispersion. However, Chen et al. (2014) applied an empirical Bayes approach where information is shared between genes (Robinson et al., 2010; McCarthy et al., 2012) and this leads to better estimates for dispersion parameters. The edgeR package consists three different dispersion estimates: common, trended and tagwise using different variations of APL function. See (Chen et al., 2014) for more details.

#### 4.2.4 Quantile transformation

It is known that genes are highly correlated which means there is a strong dependence between genes. Hence, we use quantile transformation to incorporate the dependence between genes. This approach uses the following proposition from the probability theory (Lange, 2010; p 432):

**Proposition 4.1** *Let  $X$  be a random variable with distribution function  $F(x)$ .*

- (i) *If  $F(x)$  is continuous then  $U=F(X)$  is uniformly distributed on  $[0,1]$ .*
- (ii) *Even if  $F(x)$  is not continuous, the inequality  $\Pr(F(X) \leq t) \leq t$  is still true for all  $t \in [0,1]$ .*
- (iii) *If  $F^{-1}(y) = \inf\{x: F(x) \geq y\}$  for any  $0 < y < 1$ , and if  $U$  is uniform on  $[0,1]$ , then  $F^{-1}(U)$  has distribution function  $F(x)$ .*

With the help of Proposition 4.1, we propose the following process to generate  $\mathbf{X}^{(k)}$ :

1. Assume that  $Z^{(k)} \sim MVN(\mathbf{0}, \Sigma_k)$  where  $Z_{ij} \sim N(0,1)$ .
2. If  $\Phi$  is the standard normal distribution function and  $F_k$  is the  $NB(\mu_{ij}^{(k)}, \Phi_i^{(k)})$  distribution function then the  $i$ th component of the transformed random variable

$$X_{ij}^{(k)} = F_k^{-1} \left\{ \Phi \left( Z_{ij}^{(k)} \right) \right\}, \quad (4.4)$$

has the distribution function  $F_k(x)$  which is negative binomial with parameters  $\mu_{ij}^{(k)}$  and  $\Phi_i^{(k)}$ .

We call this process as quantile transformation. Due to the discreteness of the

negative binomial random variable there may exist ambiguity while calculating the inverse probabilities. In order to remove this ambiguity of  $F_k^{-1}(x)$  we apply the following equation:

$$F_k^{-1}(q) = \inf\{x: F_k^{-1}(x)F_k(x) > q\}, q \in [0,1]. \quad (4.5)$$

In our model, inverse of the quantile transformation is applied. More precisely, we transform marginally negative binomial distribution into the underlying MVN distribution with mean  $\mathbf{0}$  and covariance matrix  $\Sigma_k$ . One can observe that dependence structure between the components of  $\mathbf{X}^{(k)}$  is involved in the model with  $\Sigma_k$  after the transformation. One can also observe that each class has a different covariance matrix that may result in better predictions and hence high performance in classification. Having a different covariance matrix for each class is also suggested by Sun and Zhao (2015) due to the fact that different disease type leads to “rewiring” of genetic networks.

#### 4.2.5 Parameter Estimation II

After applying inverse of the quantile transformation we obtain a new data matrix  $\mathbf{Z}^{(k)}$  where  $\mathbf{Z}^{(k)} \sim \text{MVN}(\mathbf{0}, \Sigma_k)$  is  $j$ th column of the new data matrix  $\mathbf{Z}^{(k)}$ . Therefore, we need to estimate the covariance matrices  $\Sigma_k$  for each class so that we can apply QDA. For that purpose, we use the simple Standard covariance matrix estimate method which is given by

$$\hat{\Sigma}_k = \frac{1}{n-1} \sum_{j=1}^n \{\mathbf{z}_j^{(k)} - \bar{\mathbf{z}}^{(k)}\} \{\mathbf{z}_j^{(k)} - \bar{\mathbf{z}}^{(k)}\}^T \quad (4.6)$$

where  $\bar{\mathbf{z}}^{(k)} = \sum_{j=1}^n \mathbf{z}_j^{(k)} / n$ . However, this simple standard estimate is not appropriate in our model due to the high dimensional RNA-Seq data. In other words, if we have the number of genes greater than or equal to the number of samples, which is highly likely to occur in RNA-Seq data, then the standard covariance matrix estimate is neither invertible nor positive definite (Tong et al., 2014).

To cope with this problem, we regularize the estimated covariance matrix using a powerful regularization technique developed by Schafer and Strimmer (2005) and Opgen-Rhein and Strimmer (2007). We implement their technique via corpcor package in R (Strimmer, 2008). Note that the regularization method developed in (Strimmer, 2005) also improves the estimate of covariance matrix and affects the classification of samples at gene expression level. The main idea in corpcor method is to shrink empirical correlation estimates  $\hat{\rho}_{ij}$  towards zero and the empirical variance estimates  $\hat{s}_i$  towards their median in order to obtain the improved estimates  $(\tilde{\rho}_{ij}, \tilde{s}_i)$  of the corresponding correlation and covariance matrices:

$$\begin{aligned}\tilde{\rho}_{ij} &= (1 - \lambda) \hat{\rho}_{ij} \\ \tilde{s}_i &= \lambda_2 s_{\text{median}} + (1 - \lambda_2) \hat{s}_i\end{aligned}$$

where the corresponding shrinkage intensities are estimated using

$$\hat{\lambda}_1 = \frac{\sum_{i \neq j} \widehat{\text{Var}}(\hat{\rho}_{ij})}{\sum_{i \neq j} \hat{\rho}_{ij}^2} \quad \text{and} \quad \hat{\lambda}_2 = \frac{\sum_{i=1}^p \widehat{\text{Var}}(s_i)}{\sum_{i=1}^p (s_i - s_{\text{median}})^2}$$

where  $s_{\text{median}}$  is the median of the empirical variances. Once the shrinkage is applied, the regularized covariance matrix estimate turns out to be positive definite. In other words, all eigenvalues are different from zero and well-conditioned (invertible). Since the shrinkage intensity estimates are evaluated analytically, it is computationally very fast and does not require any tuning parameters. To the best of our knowledge corpcor method has not been used for the classification of RNA-Seq data up to now.

#### 4.2.6 Classification

Let  $\mathbf{x}^* = [x_1^*, \dots, x_p^*]^T$  be a new observation where  $x_1^*, \dots, x_p^*$  are the components of the observed sample and  $Y^*$ , where  $Y^* \in \{1, \dots, K\}$  be the unknown class label. Before proceeding to the classification step we need to quantile transform the new observation. For this purpose, we apply inverse of the quantile transfor-

mation to the each component of  $\mathbf{x}^*$  in order to get a new vector  $\mathbf{z}^{*(k)}$  which is given as follows:

$$z_i^{*(k)} = \Phi^{-1}\{H_k(x_i^*)\}. \quad (4.7)$$

where  $z_i^*$  is  $i$ th component of  $\mathbf{z}^{*(k)}$  and  $H_k$  is a continuity-corrected version of  $F_k$ , which is defined by

$$H_k(x_i^*) = \Pr(X < x_i^*) + 0.5 \times \Pr(X = x_i^*),$$

where  $X \sim \text{NB}(\mu_{ij}^{(k)}, \Phi_i^{(k)})$  is a negative binomial random variable (Routledge, 1994). Here  $\mathbf{z}^{*(k)}$ , which represents the transformed vector of the new observation  $\mathbf{x}^*$  for class  $k$  is a new variable from the  $\text{MVN}(\mathbf{0}, \Sigma_k)$ . We use the “zscoreNBinom” function from the edgeR package in order to transform  $x_i^*$  to  $z_i^{*(k)}$ .

After the transformation process, we apply the quadratic discriminant analysis explained in Section 2.6 to classify the new observation  $\mathbf{x}^*$ . By Bayes theorem, the posterior probability of the new observation belonging to the  $k$ th class is given by

$$\Pr(Y^* = k | \mathbf{x}^*) \propto f_k(\mathbf{z}^{*(k)}) \pi_k, \quad (4.8)$$

where  $\pi_k$  is the prior probability estimated from the training set as follows:

$$\hat{\pi} = \sum_{j=1}^n \frac{I_{\{Y_j=k\}}}{n},$$

where  $I$  is the indicator function that defines the class of each observation,  $n$  is the total number of samples in all classes,

and  $f_k$  is the density

$$f_k(\mathbf{u}) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2} \mathbf{u}^T \Sigma_k^{-1} \mathbf{u}\right\},$$

evaluated at  $\mathbf{z}^{*(k)}$ . Then by replacing the density into (4.8) and taking the logarithm of the resulting function we obtain the following quadratic discriminant function for the proposed model:

$$\delta_k^{\text{qtQDA}}(\mathbf{x}^*) = -\frac{1}{2} \mathbf{v}_k^T \mathbf{v}_k + \log \pi_k \quad (4.9)$$

Where  $\mathbf{v}_k = \Sigma_k^{-1/2} \mathbf{z}^{*(k)}$ . We assign  $\mathbf{x}^*$  to the class which maximizes the discriminant function given in (4.9).

### 4.3 Application on Real RNA-Seq data

#### 4.3.1 Experimental Data Sets

In this section we give the details of the real RNA-Seq data which are publicly available:

- Cervical cancer data (Witten, 2010): Cervical cancer data consists of two classes: tumor and non-tumor. Each class has equal number of samples (29 samples) and each sample is composed of 714 microRNAs.
- Prostate cancer data (Kannan et al., 2011): Similar to cervical cancer data, prostate cancer data set consists of two distinct classes with 30 patients where 20 samples are cancer patients while 10 samples are benign matched controls.
- HapMap data (Montgomery, 2010; Pickrell, 2010): HapMap data which is quite different than the first two data sets is a data set used for ancestry estimation. Although HapMap data contains five different groups we will focus on just two of them which are CEU representing the UTAH residents with Northern and Western European Ancestry and YRI representing the Yoruba in Ibadan, Nigeria. Of all, 91 samples are from CEU while 89 samples from YRI

and each sample consists of 52,580 genes.

We note that the aforementioned data are widely used RNA-Seq data in the literature (Tan et al., 2014; Witten, 2011; Dong et al., 2016).

### 4.3.2 Implementation of Existing Classifiers

In this section we compare the performance of qtQDA model not only with some of powerful machine learning classifiers but also with specialized RNA-Seq classifiers that are given as follows:

- SVM (e1071)
- kNN (e1071)
- Logistic regression (glmnet)
- PLDA (PoiClaClu)
- NBLDA (Dong et al., 2016)
- voomDLDA (MLSeq)
- SQDA (SQDA)

We note here that the information given in brackets are corresponding R packages of classifiers used for our analysis.

We now give the implementation details of these classifiers. Let us start with machine learning classifiers. We used the same R package **e1071** for both SVM where kernel is chosen as radial basis and kNN where  $k = 1, 3, 5$ . Like NBLDA classifier (Dong et al., 2016), both classifiers were applied on log-transformed counts. The reason for doing this is that, in real data sets, the number of genes is very large and gene expression levels may show enormously different distributions (Dong et al., 2016). For logistic regression we used the **glmnet** package available in R, which uses the GLMnet method developed by Friedman (2010). For PLDA, NBLDA and voomDLDA, RNA-Seq classifiers, we used “deseq” normalization. For the last RNA-Seq classifier, SQDA we used SQDA package after log-transforming the counts.



### 4.3.3 Evaluation of the performance of classifiers

In order to compare the proposed model with existing models we need to evaluate the performance of each classification, roughly speaking we need to measure how close the predictions of the newly observed samples to the true class label of those samples. Indeed, we need to estimate the (true) error rate which can also be seen as Classification Error Rate (CER) in the literature. Once misclassification error rates are calculated we can then decide which model generates the best results for any given data set.

There exist many methods such as cross-validation, probability inequalities and bootstrap to estimate the misclassification error rate (Wasserman, 2010; Efron, 1983). In this thesis, we apply bootstrapping method to estimate the MER, the error rate where we observe false classifications, due to its simplicity and promising error rate estimate (Efron, 1983).

We now explain the bootstrapping procedure we follow. First of all, we randomly divide data into two sets: training set consisting of 70% of the data; test set consisting of 30% of the data. We then train the model using the training set. Finally, we test the model using the test set and compute the MER. To calculate the MER, we use confusion matrix given in Table 4.1:

Table 4.1. Confusion matrix

Predicted class	Actual Class		
	Positive	Negative	Total
Positive	TP	FP/Type I error	TP+FP
Negative	FN/Type II error	TN	FN+TN
Total	TP+FN	FP+TN	$N_{\text{test}}$

\*TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative,  $N_{\text{test}}$  : number of samples in the test set.

CER is computed as follows:

$$\text{CER} = \frac{\text{FP} + \text{FN}}{N_{\text{test}}},$$

which is equivalent to say that the number of misclassified samples is divided by the number of samples in the test set. We repeat the whole process 1000 times and average the CER's obtained from each iteration in order to estimate the true MER of the classification model for different number of genes.

#### **4.4 Results**

In this section, we compared and analyzed the results. We applied the proposed method on three well-known RNA-Seq data sets and compared the proposed method with the classifiers given in Section 4.3.2. Using the procedure detailed in section 4.2.2 we selected  $p = 100, 200, 300, 500, 700$  DE genes for which the error rates are estimated. We implemented other methods as recommended in their R documentations with the tuning parameters explained in Section 4.3.2.

The comparison results and the minimum error rates are given in Figure 4.1 and Table 4.2, respectively. It is obvious to see that qtQDA achieves the lowest error rate when the number of genes is 200 and 100 for the cervical cancer data and prostate cancer data, respectively. See Table 4.2. Interestingly, for the cervical cancer data, qtQDA outperforms all the other classifiers we compare with regardless of the number of genes. See Figure 4.1. For the HapMap data, qtQDA is comparable to SVM,  $k$ NN and logistic regression classifiers. On the other hand, we point out that if we are working on genomic data or medical based data instead of applying SVM or  $k$ NN we would prefer qtQDA or logistic regression classifiers as they assign samples to one of the classes with a probability score. The probability score may play an important role for further diagnostic procedures that can be associated with different risks.

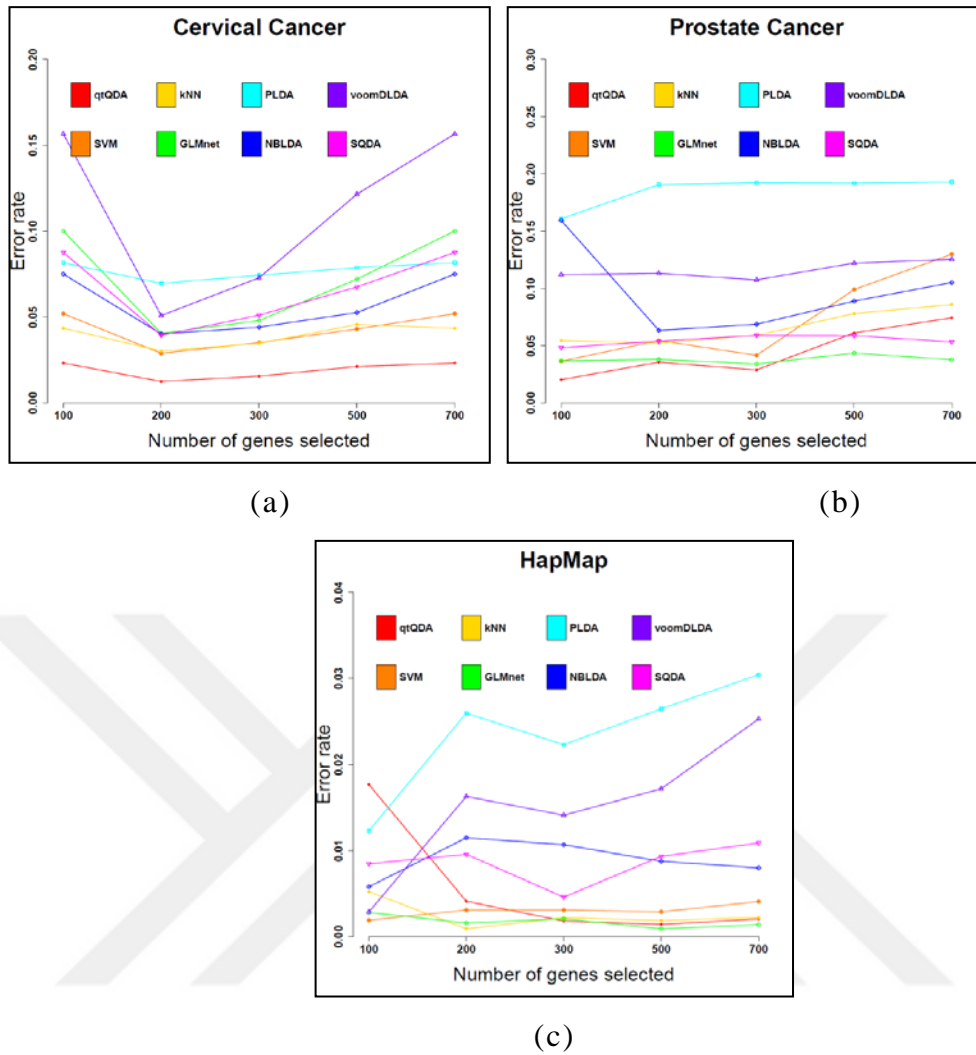


Figure 4.1. Classification error rate as a function of the number of genes chosen for classification of the (a) Cervical Cancer, (b) Prostate Cancer and (c) HapMap data sets

Table 4.2. Minimum error rate achieved for each classifier in each data set.

Method	Cervical Cancer	Prostate Cancer	HapMap
qtQDA	0.0125 (200)	0.0203 (100)	0.0018 (300)
SVM	0.0276 (100)	0.0364 (100)	0.0014 (500)
kNN	0.0277 (100)	0.0523 (200)	0.0009 (200)
GLMnet	0.0406 (200)	0.0341 (300)	0.0009 (500)
PLDA	0.0608 (100)	0.1609 (100)	0.0123 (100)
NBLDA	0.0402 (200)	0.0634 (200)	0.0058 (100)
voomDLDA	0.0425 (100)	0.1076 (300)	0.0029 (100)
SQDA	0.0318 (100)	0.0483 (100)	0.0046 (300)

\* The number of genes used to obtain the minimum error rate is reported in brackets.

## 4.5 Discussion

In this thesis, we proposed a new classifier qtQDA using RNA-Seq expression profiles which incorporate dependence structure between genes. The proposed model basically integrates quantile transformation with quadratic discriminant analysis for the RNA-Seq data classification. Therefore, instead of applying log-transformation to the counts we use quantile transformation and then apply quadratic discriminant analysis where we have class specific covariance matrices.

Although it has been shown that classifiers making unrealistic assumptions, i.e. assuming that genes are independent, can perform well for microarray data sets (Dudoit et al., 2002) our results point out the significance of the dependence structure between genes for RNA-Seq data classification problems.

The proposed model has two fundamental advantages over existing classifiers. The first advantage is that the model does not include any tuning parameters that have to be determined by cross-validation. This simplifies the application of the algorithm in practice. The second advantage is that it is computationally faster than the recently proposed RNA-Seq classifiers; SQDA method (Sun and Zhao, 2015) and the Copula method (Zhang, 2017) which incorporate the dependency between genes into the model. Since the method proposed by Zhang (2017) has no publicly available package or the algorithm we could not implement their classification approach. However, it is stated in the paper that this Gaussian copula based classifier uses a complicated Bayesian approach in combination with Metropolis-Hasting algorithm and Gibbs sampling. Thus, Zhang (2017) acknowledges that the required time for the computations is time consuming even in C++ programming language which is known as one of the fast programming languages.

Moreover, we focused on classifying samples at gene expression levels and we evaluated the classification performance only in terms of error rate not the sparsity, which is the number of genes used in classification process. We applied edgeR pipeline for gene selection which simply selects informative genes for

distinguishing between classes. On the other hand, developing a sparse version of the proposed classifier, qtQDA identifying less informative genes and reducing their impact in the classification model to zero can be further studied. A sparse version of qtQDA may lead to more efficient classification and high accuracy in the performance of the classification.



## **CHAPTER 5: A NEW LOCAL COVARIANCE MATRIX ESTIMATION FOR qtQDA**

Dependence relation between random variables in a data set is one of the broadly studied topic in statistical data analysis particularly in data classification as it can effect the performance of the classification model. The dependence relation can be incorporated either by covariance matrices or copula functions. In this thesis, we focus on the true estimation of covariance matrices which are used in the classification model. The simplest way of estimating the covariance matrix is to use Maximum Likelihood Estimator (MLE). However, since the medical or biological data sets contain highly correlated variables, the sample covariance matrix estimated by MLE may not reflect the true dependence relation between variables and this may lead to false inferences.

There exist a few approaches proposed recently in order to improve the covariance matrix estimation (Matteoli et al., 2010; Velasco-Forero et al., 2015). Caefer and Rotman (2009), for instance, proposed a quasi-local estimation approach for the covariance matrix which is estimated locally. They define dependence regions where variance of neighbours surrounding the reference point is used. Like Caefer and Rotman's approach, Oruc and Ucer (2009) constructed local dependence map with the help of a new methodology called local dependence function. This new approach has the capability of identifying three regions which are positive, negative and zero dependence. Application of the new approach on real medical data sets has shown that dependence structure based on local dependence functions is more informative.

Since it is known that RNA-Seq data sets consist of a large number of genes, the dependence structure of those genes is critical and important for classification of new samples. On the other hand, a new observation might have an individual impact on the estimation of the covariance matrix which may lead to a better classification performance. Therefore, in this part of the thesis, we propose a new approach for covariance matrix estimation which can be applied in qtQDA. We call this new approach local covariance matrix estimate. We have shown that implementing local covariance matrix in qtQDA model increases classification

accuracy on real gene expression data sets. We note here that, the local covariance matrix is estimated for each new sample. Thus, qtQDA classifier turns to an adaptive algorithm and we redefine qtQDA as L-qtQDA, i.e. Local-quantile transformed Quadratic Discriminant Analysis.

## ***5.1 Local Dependence Functions for Multivariate Normal Distributions***

In this section, we give some preliminaries/definitions about local dependence functions.

### **5.1.1 The Local Dependence Function**

Let  $F(x, y)$  be the joint cumulative distribution function and  $f(x, y)$  be the joint probability density function of a continuous bivariate random variable  $(X, Y)$ . Assume that  $F_X(x)$ ,  $f_X(x)$  are marginal and probability density functions of  $X$ ;  $F_Y(y)$ ,  $f_Y(y)$  are marginal and probability density functions of  $Y$ . Then given a pair of random variables  $(X, Y)$ , correlation coefficient between  $X, Y$  is given as

$$\rho(X, Y) = \frac{E(X-EX)(Y-EY)}{\sqrt{E(X-EX)^2}\sqrt{E(Y-EY)^2}} \quad (5.1)$$

Basically, Pearson correlation coefficient measures the linear dependence between the pair of random variables  $(X, Y)$  and it is also called the measure association (Bairamov and Kotz, 2000; Bairamov et al., 2003). However, the level of association may vary locally. In order to measure the local dependency between random variables  $X$  and  $Y$ , Bairamov and Kotz (2000) defined a new local dependence function as follows:

$$H(x, y) = \frac{E(X-E(X|Y=y))(Y-E(Y|X=x))}{\sqrt{E(X-E(X|Y=y))^2}\sqrt{E(Y-E(Y|X=x))^2}} \quad (5.2)$$

which is derived from the Equation (5.1) with the help of replacing the expectations  $EX$  and  $EY$  by the conditional expectations  $E(X|Y = y)$  and

$E(Y|X = x)$ , respectively. The local dependence function  $H(x, y)$  represents the dependence between  $X$  and  $Y$  at the point  $(x, y)$ . It can be interpreted that the local dependence function can identify the impact of  $X$  on  $Y$  “conditionally on  $X$  and  $Y$  being in a neighbourhood of the point  $(x, y)$ ” (Bairamov and Kotz, 2000; Bairamov et al., 2003).

Let  $\varepsilon_X(y) = EX - E(X|Y = y)$  and  $\varepsilon_Y(x) = EY - E(Y|X = x)$ . After some simplifications the local dependence function can be rewritten as follows:

$$H(x, y) = \frac{\text{Cov}(X, Y) + \varepsilon_X(y)\varepsilon_Y(x)}{\sqrt{\sigma_X + \varepsilon_X^2(y)}\sqrt{\sigma_Y + \varepsilon_Y^2(x)}} \quad (5.3)$$

If we divide both numerator and denominator by  $\sigma_X\sigma_Y$  then we obtain

$$H(x, y) = \frac{\rho + \varphi_X(y)\varphi_Y(x)}{\sqrt{1 + \varphi_X^2(y)}\sqrt{1 + \varphi_Y^2(x)}} \quad (5.4)$$

where  $\rho = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y}$ ,  $\varphi_X(y) = \frac{\varepsilon_X(y)}{\sigma_X}$ ,  $\varphi_Y(x) = \frac{\varepsilon_Y(x)}{\sigma_Y}$ .

We now give the properties of local dependence function in the following lemma. See Bairamov and Kotz (2000) and Bairamov et al. (2013) for more details.

*Lemma 5.1 Let  $(X, Y)$  be a pair of random variables and  $H(X, Y)$  be the local dependence function given in (5.2). Then the local dependence function has the following properties:*

- (i) If  $X$  and  $Y$  are independent then  $H(x, y) = 0$  for any  $(x, y) \in N_{XY}$ .
- (ii)  $|H(x, y)| \leq 1$ , for all  $(x, y) \in N_{XY}$ .
- (iii) If  $|H(x, y)| = 1$  for some  $(x, y) \in N_{XY}$  then  $\rho \neq 0$ .
- (iv) Let  $E(X|Y = y)$  and  $E(Y|X = x)$  are differentiable functions. If  $|H(x, y)| = 0$  for any  $(x, y) \in N_{XY}$  then  $E(X|Y = y)$  or  $E(Y|X = x)$  or both are constant.



(v) Let  $|\rho| = 1$  and assume that  $|H(x, y)| = 1$  at a point  $(x, y)$  then  $\varepsilon_X(y) = \varepsilon_Y(x)$  up to a sign.

(vi) The point  $(x^*, y^*)$  satisfying  $\varphi_X(y^*) = \varphi_Y(x^*) = 0$  is a saddle point of  $H$  and  $H(x^*, y^*) = \rho$ .

One can estimate the local dependence function from the data at hand. Let  $(X_i, Y_i), i = 1, 2, \dots, n$  be the data set. Assuming that  $\phi$  is an integrable kernel function with short tails and  $h_n \rightarrow 0$  is a width sequence tending zero at approximate rates, Nadaraya (1964) and Watson (1964) proposed the following estimates for the regression functions  $E(X | Y = y)$  and  $E(Y | X = x)$ :

$$A_X^{(n)}(y) = \frac{\sum_{i=1}^n X_i \phi\left(\frac{y-Y_i}{h_n}\right)}{\sum_{i=1}^n \phi\left(\frac{y-Y_i}{h_n}\right)} \quad \text{and} \quad A_Y^{(n)}(x) = \frac{\sum_{i=1}^n Y_i \phi\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n \phi\left(\frac{x-X_i}{h_n}\right)}$$

Using Nadaraya and Watson's estimates, Bairamov and Kotz (2000) suggested the following estimate for the local dependence function

$$H^n(x, y) = \frac{\rho^{(n)} + \frac{(\bar{X} - A_X^{(n)}(y))(\bar{Y} - A_Y^{(n)}(x))}{s_X s_Y}}{\sqrt{1 + \frac{(\bar{X} - A_X^{(n)}(y))^2}{s_X^2}} \sqrt{1 + \frac{(\bar{Y} - A_Y^{(n)}(x))^2}{s_Y^2}}} \quad (5.5)$$

where

$\rho^{(n)}$  is an estimate for Pearson correlation coefficient,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, s_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

## 5.2 A new estimate of local dependence function

Since in qtQDA model we estimate the covariance matrix, using Bairamov and Kotz's estimate given in Equation (5.5) we suggest the following estimate for the local dependence function

$$\widehat{H}^{(n)}(x, y) = s_x s_y \widetilde{H}^{(n)}(x, y) \quad (5.6)$$

Now, it turns to figure out the optimal kernel function and the optimal bandwidth which improves the true covariance matrix estimate. It is given in (Silverman, 1986) that the optimal choice for  $h$  (i.e., the bandwidth that minimises the mean integrated squared error) is

$$h_n = \left(\frac{4\widehat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06\widehat{\sigma}n^{-1/5} \quad (5.7)$$

where  $\widehat{\sigma}$  is the standard deviation of the samples. This approximation is called Gaussian approximation, known also as Silverman's rule of thumb (Silverman, 1986). Note here that, Silverman's rule of thumb approximation is used for downstream analysis. For the kernel function, we applied the triangular kernel function given in Equation (5.8):

$$\phi(u) = 1 - |u|, |u| \leq 1 \quad (5.8)$$

## 5.3 Results

In this section of the thesis, we compare the performance of the proposed model with sample covariance matrix and covariance matrix estimated by local dependence function on two real RNA-Seq data sets: cervical cancer and HapMap data given in Section 4.3.1. We implement the same algorithm and the same process for gene selection explained in Chapter 3.2. The whole procedure is repeated 300 times for different number of genes (20, 50, 100, 200, 300, 500) and the misclassification error rate is computed by bootstrapping method described in Section and then we average the error error rates from each iteration

to estimate the true error rate of misclassification.

The comparison results are given in Table 5.1, Figure 5.1 and Figure 5.2. We obviously see that L-qtQDA where we estimate covariance matrix by local dependence function performs generally better than qtQDA where we estimate covariance matrix by MLE for both cervical cancer data and HapMap data. Even though qtQDA achieves the lowest error rate at 200 genes for cervical cancer data L-qtQDA achieves the lowest error rate at 50 genes.

Table 5.1. Classification error rates for cervical cancer and HapMap data sets

<b>Data</b>	<b># of genes</b>	<b>qtQDA</b>	<b>L-qtQDA</b>
<b>Cervical Cancer</b>	20	<b>0.0367</b>	0.0372
	50	0.0280	<b>0.0265</b>
	100	0.0126	<b>0.0124</b>
	200	<b>0.0117</b>	0.0122
	300	0.0161	<b>0.0159</b>
	500	0.0189	<b>0.0170</b>
<b>HapMap</b>	20	0.0172	<b>0.0166</b>
	50	0.0064	<b>0.0057</b>
	100	0.0448	<b>0.0434</b>
	200	<b>0.0120</b>	0.0116
	300	0.0074	<b>0.0073</b>
	500	<b>0.0106</b>	0.0109

#### **5.4 Discussion**

While working on cancer prediction, one of the pivotal steps is to incorporate the true/accurate covariance matrix into the classification model. This chapter covers a new approach for estimating the class-specific covariance matrices by using local dependence function to be used in RNA-Seq data classification and the impact of differently estimated covariance matrices on RNA-Seq data classification. Assuming that the dependencies between genes are locally defined rather than complete dependency, this study illustrates that the effectiveness of

locally estimated covariance is higher than simple covariance matrix on real RNA-Seq data classification.

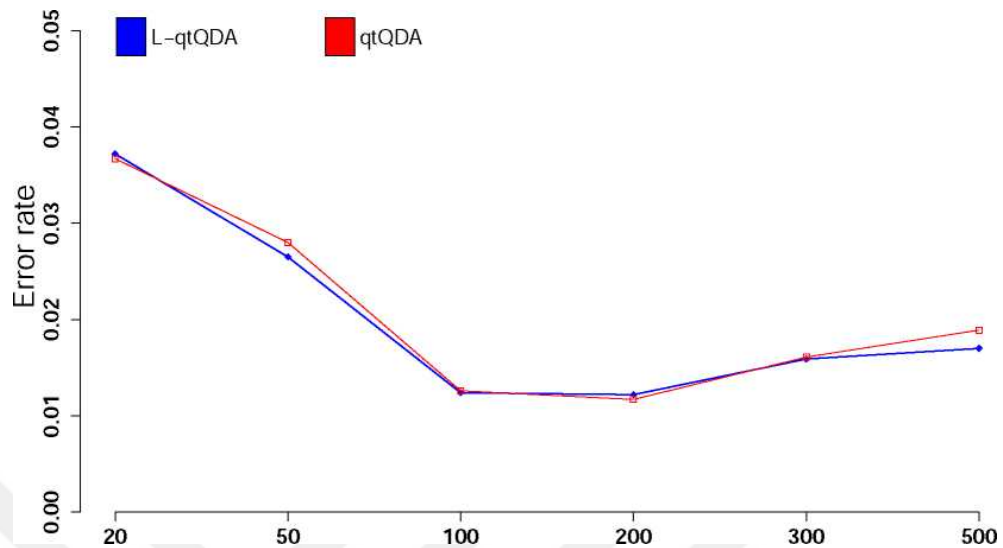


Figure 5.1. Classification error rate as a function of the number of genes chosen for classification of the cervical cancer

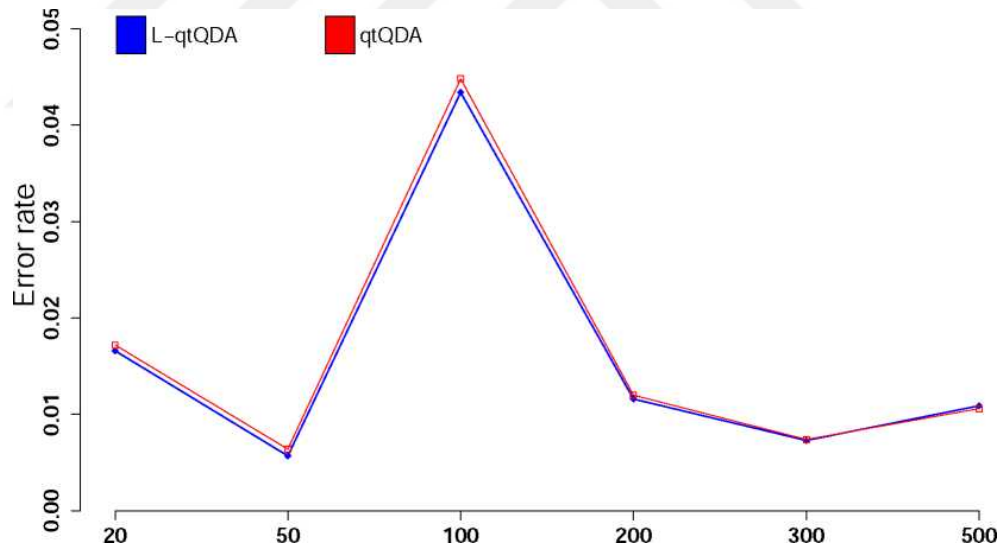


Figure 5.2. Classification error rate as a function of the number of genes chosen for classification of the HapMap data

Underlining that this thesis solely utilizes triangular kernel function and Gaussian bandwidth in local dependency calculation, we note that implementation of different kernel functions and a different optimal bandwidth selection may improve classification performances. The only disadvantage of the L-qtQDA is that the algorithm is computationally intensive due to the estimation of the local

covariance matrices. Nevertheless, we believe that this new estimation technique will be useful for classification of RNA-Seq profiles or other genomic studies.



## CHAPTER 6: CONCLUSION AND FURTHER STUDIES

In this thesis, we proposed a new classifier, quantile transformed Quadratic Discriminant Analysis (qtQDA), using RNA-Seq expression profiles which incorporate dependence structure between genes. The proposed model basically integrates quantile transformation with quadratic discriminant analysis for the RNA-Seq data classification. Therefore, instead of applying log-transformation to the counts we use quantile transformation and then apply quadratic discriminant analysis where we have class specific covariance matrices. While we quantile transform the data, which is assumed to be marginally negative binomial but dependent, we used the sophisticated edgeR methodology for the parameter estimation of negative binomial marginals. To the best of our knowledge, edgeR has only been used for discovering the differentially expressed genes. Thus, we appear to be the first to use edgeR methodology directly at the classification stage of RNA-Seq classification problems. For the estimation of class specific covariance matrices we used two different techniques: Maximum Likelihood Estimator (MLE) and local dependence function estimator. We proposed a new covariance estimator based on local dependence function and compare it with MLE. In either case a powerful regularization technique which can be applied by corpcor package and has never been used for RNA-Seq data classification before is applied so that QDA can be performed on the quantile transformed data.

In order to increase the performance of the classification one of the crucial steps in the proposed algorithm is the gene selection step. The gene selection approach we implemented is obviously a very simple approach to select genes which are informative for distinguishing between classes. Hence, developing a sparse version of qtQDA, which can detect less informative genes and reduce their effect to zero can be further investigated. We expect even a better performance in real RNA- Seq data classification.

Dependency between genes can be modelled either with copula functions or multivariate distribution functions. In this thesis, we incorporated the dependency between genes into the model with the help of Multivariate Normal

distribution. Multivariate normal distribution does not only simplifies the algorithm but also enables us to implement Gaussian QDA. On the other hand, it is known that copula functions are powerful functions which describe the dependency between variables. Although Zhang (2016) developed a new RNA-Seq classifier via Gaussian copula functions there is no publicly available package or the algorithm that we could implement the classification approach described in (Zhang, 2016). Thus, in the future study, improving a classifier via different copula functions (including Gaussian copula functions) will be explored. The impact of different copula functions can be compared and analyzed. Additionally, an R package for Copula based RNA-Seq data classification can be further written and published.

Last but not least the proposed classification algorithm can be extended for multi-class classification problems and can then be compared with the existing classifiers.

## REFERENCES

AbuElQumsan M. (2018) *Assessment of Supervised Classification Methods for the Analysis of RNA-seq Data*. Doctoral Thesis. Aix-Marseille Universite, Faculte des sciences de Luminy, available at: <https://www.theses.fr> (Accessed: 18 July 2019).

Altman, N.S. (1992) “An introduction to kernel and nearestneighbor nonparametric regression”. *The American Statistician*, Vol. 46(3), pp. 175-185.

Anders, S. and Huber, W. (2010) *Differential expression analysis for sequence count data*. *Genome Biol.*, Vol. 11, pp. 1-12.

Bairamov, I. and Kotz, S. (2000) *On local dependence function for multivariate distributions*, *New Trends in Probability and Statistics*, Vol. 5, pp. 27-44.

Bairamov, I., S. Kotz. and Kozubowski, T.J. (2003) *A new measure on linear local dependence*, *Statistics*, Vol. 37, pp. 243-258.

Bullard, J.H., Purdom, E., Hansen, K.D. and Dudoit, S. (2010) *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*, *BMC Bioinformatics*, pp. 11- 94.

Caefer C.E. and Rotman, S.R. (2009) *Local covariance matrices for improved target detection performance*, in: *Proceedings of the 1st Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS '09)*, pp. 1-4.

Chen, Y., Lun, A.T. and Smyth, G.K. (2014) *Differential expression analysis of complex RNA-seq experiments using edgeR*, in Datta S., Nettleton D., eds., *Statistical analysis of next generation sequencing data*. *Frontiers in Probability and the Statistical Sciences*. New York: Springer, Cham, pp. 51-74.



Cox, D R. and Reid, N. (1987) *Parameter orthogonality and approximate conditional inference*, J. R. Stat. Soc. Series B, Vol. 49, pp. 1-39.

Dong, K., Zhao, H., Tong, T. and Wan, X. (2016) *NBLDA: negative binomial linear discriminant analysis for RNA-Seq data*, BMC Bioinformatics, Vol. 17(369), pp. 1-10.

Duda, R. O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*, 2nd ed. Wiley Interscience.

Efron, B. (1983) *Estimating the Error Rate of a Prediction Rule*, Journal of the American Statistical Association, Vol. 78, pp. 316-333.

Fang Z., Martin, J. and Z. Wang. (2012) *Statistical methods for identifying differentially expressed genes in RNA-Seq experiments*, Cell & Bioscience, Vol. 2(26), pp. 1-8.

Fisher, R.A. (1936) *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics, Vol. 7(2), pp. 179-188.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) *Regularization paths for generalized linear models via coordinate descent*, Journal of statistical software, Vol. 33(1), pp. 1-22.

Fu, X., Fu, N., Guo, S., Yan, Z., Xu, Y., Hu, H., Menzel, C., Chen, W., Li, Y. and Zeng, R. (2009) *Estimating accuracy of RNA-seq and microarrays with proteomics*, BMC Genom., Vol. 10(161), pp. 1-9.

Haas, B.J. and Zody, M.J. (2010) *Advancing RNA-seq analysis*, Nat. Biotech., Vol. 28(5), pp. 421-423.

Hoerl, A.E. and Kennard, R.W. (1970) *Ridge regression: biased estimation for non-orthogonal problems*, Technometrics, Vol. 12(1), pp. 55-67.

Holt R.A. and Jones, S.J. (2008) *The new paradigm of flow cell sequencing*, Genome Res., Vol. 18, pp. 839-846.

Kannan, K., Wang, L., Wang, J., Ittmann, M.M., Li, W. and Yena, L. (2011) *Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing*, ed. B. Vogelstein. *Proceedings of the National Academy of Sciences*, Vol. 108(22), pp. 9172-9177.

Klaus, B. (2015) *Statistical relevance-relevant statistics, part I*, The EMBO Journal, Vol. 34(22), pp. 2727-2730.

Lange, K. (2010). *Numerical analysis for statisticians*. New York: Springer.

Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. (2014) *voom: precision weights unlock linear model analysis tools for RNA-Seq read counts*, Genome Biology, Vol. 15, pp. 1-17.

Li, J., Witten, D.M., Johnstone, I.M. and Tibshirani, R. (2012) *Normalization, testing, and false discovery rate estimation for RNA sequencing data*. Biostatistics, Vol. 13(3), pp. 523-538.

MacKenzie, R.J. (2018) DNA vs. RNA - 5 Key Differences and Comparison [Online]. Available at: <https://www.technologynetworks.com/genomics/lists/what-are-the-key-differences-between-dna-and-rna-296719> (Accessed: 15 October 2019).

Mardis, E. R. (2008) *Next-generation DNA sequencing methods*, Annu. Rev. Genomics Hum. Genet., Vol. 2849, pp. 387-402.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) *RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays*, Genome Res, Vol. 18, pp. 1509-1517.

Matteoli, S., Diani, M. and Corsini, G. (2010) *Improved estimation of*

*local background covariance matrix for anomaly detection in hyper-spectral images*, Optical Engineering, Vol. 49, pp. 1-16.

McCarthy D.J., Chen, Y. and Smyth, G.K. (2012) *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation*. Nucleic Acids Research, Vol. 40(10), pp. 4288-4297.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd edn. Chapman & Hall/CRC, Boca Raton, Florida.

Metzker, M.L. (2010) *Sequencing technologie-the next generation*, Nature Reviews Genetics, Vol. 11, pp. 31-46.

Montgomery, S.B. (2010) *Transcriptome genetics using second generation sequencing in a Caucasian population*, Nature, Vol. 464(7289), pp. 773-777.

Mortazavi, A., Williams, B.A, McCue, K., Schaeffer, L., and Wold, B. (2008) *Mapping and quantifying mammalian transcriptomes by RNA-Seq*, Nat. Methods, Vol. 5, pp. 621-628.

Nadaraya, E.A. (1964) *On estimating regression*, Theory Probab. Appl., Vol. 9, pp. 141-142.

Nelder, J.A. and Wedderburn, R.W.M. (1972) *Generalized linear models*, J. Roy. Stat. Soc. A, Vol. 135, pp. 370-384.

Opgen-Rhein, R. and Strimmer, K. (2007) *Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach*. Statistical applications in genetics and molecular biology, Vol. 6(1), pp. 1-18.

Oruc, O.E. and Ucer, B. (2009) *A new method for local dependence map and its applications*, Turkiye Klinikleri J. Biosta., Vol. 1, pp. 1-8.

Özdoğan, M. (2018) *Yeni nesil dizileme (Next-Generation Sequencing, NGS) nedir? Onkolojide kullanımı.* [Online]. Available at: <https://www.drozdogan.com/yeni-nesil-dizileme-next-generation-sequencing-ngs-nedir-kanserde/> (Accessed: 15 October 2019).

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A., Brown, P.O. and Botstein D. (2000) *Molecular portraits of human breast tumours*, Nature, Vol. 294406(6797), pp. 747-752.

Pickrell, J.K. (2010) *Understanding mechanisms underlying human gene expression variation with RNA sequencing*, Nature, Vol. 464(7289), pp. 768-772.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth G.K. (2015) *limma powers differential expression analyses for RNA-sequencing and microarray studies*, Nucleic Acids Research, Vol. 43(7), pp. 1-13.

Robinson, M.D. and Smyth, G.K. (2008) *Small-sample estimation of negative binomial dispersion with applications to SAGE data*, Biostatistics, Vol. 9, pp. 321-332.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) *edgeR: A Bioconductor package for differential expression analysis of digital gene expression data*, Bioinformatics, Vol. 26, pp. 139-140.

Routledge, R.D. (1994) *Practicing safe statistics with the mid-p*, Canadian Journal of Statistics, Vol. 29922(1), pp. 103-110.

Sayad S. (2019) *Support Vector Machine Classification (SVM)* [online]. Available at: [http://saedsayad.com/support\\_vector\\_machine.htm](http://saedsayad.com/support_vector_machine.htm) (Accessed: 12 October 2019).

Schafer, J. and Strimmer, K. (2005) *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*, Stat. Appl. Genet. Mol. Biol, Vol. 4(1), pp. 1-30.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall/CRC.

Slack, J.M.W. (2014) *Genes-A Very Short Introduction*. Oxford University Press, Oxford.

Stewart, G. (1973) *Introduction to Matrix Computations*. Computer Science and Applied Mathematics. Academic Press, NY.

Strimmer, K. (2008) *Comments on: Augmenting the bootstrap to analyze high dimensional genomic data*. Test, Vol. 17(1), pp. 25-7.

Sun J. and Zhao, H. (2015) *The application of sparse estimation of covariance matrix to quadratic discriminant analysis*, BMC Bioinformatics, Vol. 16(1), pp. 1-9.

Tan, K.M., Petersen, A. and Witten, D. (2014) *Classification of RNA-seq data*. In Datta, S. and Nettleton, D., editors. Statistical analysis of next generation sequencing data, Springer, Vol. 309, pp. 219-246.

Tibshirani, R. (1996) *Regression shrinkage and selection via the lasso*. J. Roy. Stat. Soc. Ser. B (Methodological), Vol. 58, pp. 267-288.

Tong, T., Wang, C. and Wang, Y. (2014) *Estimation of variances and covariances for high-dimensional data: a selective review*, Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 6(4), pp. 255-64.

Vapnik, V. (2000) *The Nature of Statistical Learning Theory*. 2nd ed. New York: Springer-Verlag.

Velasco-Forero, S., Chen, M., Goh, A. and Pang, S.K. (2015) *Comparative Analysis of covariance matrix estimation for anomaly detection in hyperspectral images*, IEEE J. Sel. Topics Signal Process, Vol. 9, pp. 1061-1073.

Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010) *DEGseq: an R package for identifying differentially expressed genes from RNA-Seq data*, Bioinformatics, Vol. 26, pp. 136-138.

Wang, Z., Gerstein, M. and Snyder, M. (2009) *RNA-seq: a revolutionary tool for transcriptomics*, Nature Reviews Genetics, Vol. 10(1), pp. 57-63.

Wasserman, L. (2010) *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Watson, G.S. (1964) *Smooth regression analysis*. Sankhya A., Vol. 26, pp. 359-372.

Watson, J. and Crick, F. (1953) *A structure for deoxyribose nucleic acid*, Nature, Vol. 171, pp. 737-738.

Wilimitis D. (2018) *The Kernel Trick in Support Vector Classification* [Online]. Available at: <https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f> (Accessed: 15 September 2019).

Witten, D.M. (2011) *Classification and clustering of sequencing data using a Poisson model*, Annals Appl Stat., Vol. 5 (2), pp. 493-518.

Witten, D., Tibshirani, R., Guoping Gu, S., Fire A. and Lui, W. (2010) *Ultra- high through-put sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls*, BMC Biology, Vol. 8(58), pp. 1-14.

Yao, Z. and Ruzzo, W.L. (2006) *A Regression-based K nearest neighbor*

*algorithm for gene function prediction from heterogeneous data*, BMC Bioinformatics, Vol 7, pp. 1-11.

Yu, D., Huber, W. and Vitek, O. (2013) *Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size*, Bioinformatics, Vol. 29(10), pp. 1275-1282.

Zararsiz G. (2015) *Development and Application of Novel Machine Learning Approaches for RNA-Seq Data Classification*. Doctoral Thesis. Hacettepe University, Institute of Health Sciences, Available at: <http://openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/998?locale-attribute=en> (Accessed: 28 September 2019).

Zararsiz, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G.E., Duru, P., Unver, T. and Ozturk, A. (2017a) *A comprehensive simulation study on classification of RNA-Seq data*, Plos One, Vol. 12 (8), pp. 1-19.

Zararsiz, G., Goksuluk D., Klaus, B., Korkmaz S., Eldem V., Karabulut, E. and Ozturk, A. (2017b) *voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data*, PeerJ. Vol. 5, pp. 1-27.

Zhang, Q. (2017) *Classification of RNA-Seq data via Gaussian copulas*, The ISI's Journal for the Rapid Dissemination of Statistics Research, Vol. 6, pp. 171-183.

## VITA

Necla Koçhan was born in Izmir, Turkey, on December 16, 1987. She finished her high school education in Çağdaş Eğitim College in Izmir. In 2010 she received her bachelor's with third degree from Izmir University of Economics Department of Mathematics. At the same time she completed Department of Economics as a double major. After graduating from university she started her master in Applied Statistics in Izmir University of Economics. In 2013 she finished her thesis called Fuzzy Bayes Classification, supervised by Assist. Prof Dr. Güvenç Arslan. In 2013 she was accepted in Ph.D. programme in Applied Mathematics and Statistics in Izmir University of Economics. She has been awarded PhD scholarship by TUBITAK. She continued her research as a visiting scholar at Walter and Eliza Hall Institute from June 2018 to February 2019 and this was supported by TUBITAK as well. She worked as a research assistant in Izmir University of Economics Department of Mathematics between the years 2010 and 2017. In 2019 she completed all the requirements for the Doctor of Philosophy degree in the Graduate Program of Applied Mathematics and Statistics at Izmir University of Economics under the supervision of Prof. Dr. Gözde Yazgı Tütüncü. Her research interests include bioinformatics, statistical data analysis and computational biology.