T.C.
REPUBLIC OF TURKEY
HACETTEPE UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

COUNT BASED CLUSTERING AND CLASSIFICATION OF RNA-SEQ DATA

Dinçer GÖKSÜLÜK

Program of Biostatistics
INTEGRATED DOCTOR OF PHILOSOPHY THESIS

ANKARA
2019

T.C.
REPUBLIC OF TURKEY
HACETTEPE UNIVERSITY
GRADUATE SCHOOL HEALTH SCIENCES

COUNT BASED CLUSTERING AND CLASSIFICATION OF RNA-SEQ DATA

Dinçer GÖKSÜLÜK

Program of Biostatistics
INTEGRATED DOCTOR OF PHILOSOPHY THESIS

THESIS SUPERVISOR
Prof. Ergun KARAAĞAOĞLU

ANKARA
2019

Count Based Clustering and Classification of RNA-seq Data

Dinçer Göksülük

Supervisor: Prof. Dr. Ahmet Ergun Karaağaoğlu

This thesis study has been approved and accepted as an integrated PhD dissertation in "Biostatistics Program" by the assesment committee, whose members are listed below, on January 28, 2019.

Chairman of the Committee :	*Prof. Dr. Celal Reha ALPAR* *Hacettepe University*
Member :	*Prof. Dr. Atilla Halil ELHAN* *Ankara University*
Member :	*Prof. Dr. Erdem KARABULUT* *Hacettepe University*
Member :	*Doç. Dr. Jale Karakaya KARABULUT* *Hacettepe University*
Member :	*Doç. Dr. Derya GÖKMEN* *Ankara University*

This dissertation has been approved by the above committee in conformity to the related issues of Hacettepe University Graduate Education and Examination Regulation.

06 Subat 2019

Prof. Diclehan ORHAN, MD, PhD

Institute Manager

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

28/01/2019

Dinçer GÖKSÜLÜK

i

---

[i] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

*(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

*(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkânı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

*(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir

****** Tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

ETHICAL DECLARATION

In this thesis study, I declare that all the information and documents have been obtained in the base of the academic rules and all audio-visual and written information and results have been presented according to the rules of scientific ethics. I did not do any distortion in data set. In case of using other works, related studies have been fully cited in accordance with the scientific standards. I also declare that my thesis study is original except cited references. It was produced by myself in consultation with supervisor Prof. Dr. A. Ergun KARAAĞAOĞLU and written according to the rules of thesis writing of Hacettepe University Institute of Health Sciences.

Dinçer Göksülük

ACKNOWLEDGEMENTS

I am thankful to my supervisor, Prof. Dr. Ahmet Ergun Karaağaoğlu, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. I am also grateful to the members of the thesis examining committee, Prof. Dr. C. Reha Alpar, Prof. Dr. Erdem Karabulut, Prof. Dr. A. Halil Elhan, Assoc. Prof. Dr. Jale Karakaya Karabulut and Assoc. Prof. Dr. Derya Gökmen, whose suggestions and questions improved the clarity and quality of my thesis.

I am also thankful to Assoc. Prof. Dr. Gökmen Zararsız and Assist. Prof Dr. Selçuk Korkmaz for their helps, contributions and suggestions. They are more than friends to me, and i feel lucky to meet with them in my academic career. I would like to thank to Assist. Prof. Dr. N. Anıl Dolgun, my old roommate, for her valuable comments and contributions, and also for her great taste in music. A good song from her playlist for a good day!

I offer my thanks and regards to Dr. Duygu Aydın Haklı for her supports. I also would like to thank to colleagues in my department, our secretary Menekşe Tarla, and all of those who supported me in any respect during the completion of my thesis.

Lastly, and the most importantly, i would like to thank to my beloved family for their support during my entire life of education. I am heartily thankful to my wife who makes me feel happy all the time, for her encouragement and support.

ABSTRACT

Göksülük, D. Count based clustering and classification of RNA-Seq data. Graduate School of Health Sciences, Integrated Doctor of Philosophy Thesis in Biostatistics, Ankara, 2019. In molecular biology, gene-expression based studies are frequently used for examining transcriptional activities in different tissue samples or cell populations. Gene expression data can be used for different tasks; e.g differential expression, classification and clustering. In this thesis we focused on clustering and classification of gene expression data obtained from RNA sequencing experiment. Poisson (PLDA) and negative binomial (NBLDA) linear discriminant analyses are selected as discrete, and nearest shrunken centroids (NSC) are selected as continuous classifiers in classification part. We proposed an extension of NBLDA as sparse classifier and compared its performance with other classifiers. In clustering part, we used k-means and hierarchical clustering as continuous, and Poisson and negative binomial clustering as discrete approaches. A comprehensive simulation study is conducted for classification part under different scenarios. Furthermore, we used three different real data sets. Simulation results showed that overdispersion has an important effect on model performances. Overall, discrete models performed better in classification. Among discrete classifiers, NBLDA outperformed PLDA when data set is highly overdispersed. Moreover, our proposed algorithm performed better than NBLDA algorithm in terms of prediction accuracy and sparsity. We also applied the same classifiers to three real data sets and found that results agree with the simulation results. Clustering, on the other hand, is applied to real data sets only. Unlike classification, discrete and continuous clustering approaches performed similar on two real data sets. We did not perform a simulation study for clustering scenarios due to several reasons: (i) simulations were computationally intensive, and (ii) dissimilarity matrices cannot be calculated when data set had several thousands of features. Therefore, we were not able to generalize clustering results. In conclusion, discrete statistical approaches should be preferred for classification while discrete or continuous approaches can be preferred for clustering purpose. However, if visualizing data is of interest in clustering, data should be transformed for better graphical results.

Key Words: RNA sequencing, classification, clustering, negative binomial, Poisson, linear discriminant analysis.

ÖZET

Göksülük, D. RNA dizileme verilerinin kesikli yöntemler ile sınıflandırılması ve kümelendirilmesi. Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı Bütünleşik Doktora Tezi, Ankara, 2019. Gen ifade çalışmaları sıklıkla farklı dokularda ve hücre yapılarında genlerin aktivasyon düzeylerini ölçmek amacıyla kullanılmaktadır. Gen ifade verileri gen ekspresyonu, sınıflama ve kümeleme gibi farklı amaçlar için kullanılabilir. Bu tez kapsamında RNA dizilemeden elde edilen gen ifade verilerinin kümelenmesi ve sınıflanması üzerinde durulmuştur. Poisson (PDAA) ve negatif binom (NBDAA) doğrusal ayırma analizleri kesikli, en yakın küçültülmüş küme merkezleri (KKM) ise sürekli yöntemler olarak seçilmiştir. Ayrıca, NBDAA yönteminin bir uzantısı olarak seyrek NBDAA algoritması bu tez kapsamında geliştirilmiştir. Kümeleme analizinde ise k-en yakın kümeler ve hiyerarşik kümeleme algoritmaları sürekli, Poisson ve negatif binom kümeleme algoritmaları ise kesikli algoritmalar olarak seçilmiştir. Sınıflama analizi için farklı senaryolar altında kapsamlı bir benzetim çalışması yapılmıştır. Ayrıca, üç farklı gerçek veri seti kullanılmıştır. Benzetim çalışması aşırı yayılım parametresinin performanslar üzerinde önemli bir etkisi olduğunu göstermiştir. Genel olarak kesikli dağılımlar daha iyi sınıflama performansı göstermiştir. NBDAA yöntemi aşırı yaygın veri setlerinde PDAA yöntemine göre daha iyi performans göstermiştir. Geliştirdiğimiz seyrek NBDAA yöntemi ise NBDAA yöntemine göre sınıflama performansı ve modeldeki değişken sayısı bakımından daha iyi performans göstermiştir. Aynı sınıflama algoritmaları gerçek veri setlerine de uygulanmış ve benetim çalışmasını destekleyici sonuçlar elde edilmiştir. Kümeleme performansları iki gerçek veri setinde kesikli ve sürekli dağılımlar için benzer sonuçlar vermiştir. Kümeleme analizi için hesaplama sürelerinin çok yüksek olması ve on binlerce değişken içeren veri setlerinde uzaklık matrislerinin hesaplanamaması gibi sebeplerden dolayı benzetim çalışması yapılamamıştır. Bu nedenle kümeleme analizi sonuçları için bir genelleme yapılamamıştır. Sonuç olarak, kesikli dağılımlara dayalı yaklaşımlar RNA dizileme verilerinin sınıflamasında öncelikli olarak tercih edilmelidir. Kümeleme analizinde ise kesikli veya sürekli dağılımlar isteğe göre tercih edilebilir. Ancak, kümeleme analizinde verilerin görselleştirilmesi amaçlanıyor ise dönüşüm uygulanması daha iyi grafiksel sonuçlar elde edilmesini sağlayabilir.

Anahtar Kelimeler: RNA dizileme, sınıflama, kümeleme, negatif binom, Poisson, doğrusal ayırma analizi.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ALL	Acute lymphocytic leukemia
AML	Acute myeloid leukemia
cDNA	Complementary DNA
DE	Differential expression
DNA	Deoxyribonucleic acid
DLDA	Diagonal linear discriminant analysis
DQDA	Diagonal quadratic discriminant analysis
GLM	Generalized linear model
KICH	Kidney chromophobe carcinoma
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LDA	Linear discriminant analysis
limma	Linear models for microarray data
lncRNA	Long non-coding RNA
LOWESS	Locally weighted scatter plot smoothing
miRNA	Micro RNA
mRNA	Messenger RNA
MLE	Maximum likelihood estimation
NBLDA	Negative binomial linear discriminant analysis
NGS	Next generation sequencing
NSC	Nearest shrunken centroids
PCC	Principal component classification
PLDA	Poisson linear discriminant analysis
RF	Random forests
rlog	Regularized logarithmic transformation
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
RPKM	Reads per kilobase per million mapped reads
sPLDA	Sparse Poisson linear discriminant analysis
QDA	Quadratic discriminant analysis
TMM	Trimmed mean of M-values
voom	Variance modeling at the observational level
voomNSC	Voom-based nearest shrunken centroids
vst	Variance stabilizing transformation

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In molecular biology, gene-expression based studies have great importance on examining the transcriptional activities in different tissue samples or cell populations (1). During the last two decades, a vast of the literature has worked on evaluating the effect of transcriptional expression patterns on specific conditions such as biological conditions, tumour subtypes, etc (2–4). These transcripts may refer to genes, protein coding sequences, micro RNAs (miRNA), long non-coding RNAs (lncRNA), etc. For simplicity of language, we will use the term 'gene' throughout this thesis.

With the recent advances, it is now feasible to examine the expression levels of thousands of genes at the same time. This leads researchers to focus on multiple analysis tasks: (i) class discovery, (ii) class comparison and (iii) class prediction. Class discovery is used when the outcome of interest is unknown. The aim of this analysis is to discover clusters to get information from data, where within-cluster observations are similar and between-cluster observations are dissimilar to each other. In such way, researchers are able to identify the disease subtypes based on the gene-expression profiles of observations. Class comparison is perhaps the most widely used analysis task in gene expression analysis. The objective is to compare the expression profiles of genes among interested class conditions. In this way, researchers identify probably large number of significant genes that have an effect on these conditions. In class prediction, the outcome of interest is known. The aim is to assign observations to already known class conditions based on gene expression profiles for future predictions. In this wise, a small subset of genes, that are relevant with the condition, are identified and the data is trained to obtain learning rules for prediction. After the training process, new test observation classes are predicted based on the training information. This gene expression based prediction is very useful for molecular diagnosis of diseases. In statistical terminology, 'clustering analysis' term is used for class discovery; 'differential expression (DE) analysis' term is used for class comparison and 'classification analysis' term is used for class prediction problems (2, 5).

Microarray and next-generation sequencing (NGS) technologies are the recent high-throughput technologies for quantifying gene expression. RNA sequencing (RNA-Seq) is the technique which uses the capabilities of NGS technology to characterize and quantify gene expression (6). Although both microarray and RNA-Seq techniques provide the expression levels of thousands of genes simultaneously, RNA-Seq has become the state of the art approach due to its major advantages. Firstly, microarray technique provides noisy data because of cross-

hybridization. In order to obtain a clean gene-expression data, researchers start with filtering hundreds or thousands of non-informative genes (7). Secondly, researchers can only work with the genes which are present on the array. With the use of RNA-Seq technique, researchers are able to work with more accurate data with less information loss and are able to detect novel transcripts and isoforms (8).

A great deal of algorithms are developed or adapted for microarray based gene expression studies. Classical statistical algorithms have not been directly applied due to the nature of gene expression data. In a typical microarray data, number of observations is mostly expressed in tens or hundreds, while the number of genes is usually in thousands. For this reason, many approaches are proposed to estimate the true gene-level variance for small sample size setting, and to adjust significance values to overcome the multiple testing problems. Linear models for microarrays and RNA-Seq data and significance analysis of microarrays (SAM) are modified version of the t-test and among the most powerful approaches for DE analysis (9, 10). Linear models for microarray data (limma) use an empirical Bayesian approach and linear models to analyse microarray experiments (9). SAM is similar to limma method, however has some differences such as using permutation tests and nonparametric statistics (data may not follow a normal distribution) (10). To cluster microarray data, Monti et al. (11) developed consensus clustering algorithm which is based on resampling techniques and stability of obtained clusters. Dudoit and Fridlyand (12) proposed Clest algorithm which initially splits the data into training and test sets, and then clusters the training data. Next, this algorithm builds a classifier from cluster labels and predicts the class labels of test data. Finally, a similarity score is calculated from clustering and prediction results of test data. Dudoit and Fridlyand (13) also proposed bagged clustering method for microarray data clustering. A large number of studies are applied for classification of microarray data. Díaz-Uriarte and Alvarez de Andrés (5) applied random forests (RF) method; Brown et al. (14) applied support vector machines (SVM) classifier for microarray classification. Both studies showed that these two machine learning approaches are applicable for high-dimensional microarray datasets. Dudoit et al. (15) extended linear and quadratic discriminant analysis (LDA, QDA) algorithms for this purpose. The authors assumed the genes are independent each other, used diagonal covariance matrices in calculating discriminant scores and named these two algorithms as diagonal linear and diagonal quadratic discriminant analysis (DLDA, DQDA). Both algorithms are applicable in high-dimensional settings. However, very com-

plex discriminant models are obtained with the increasing number of genes. To overcome this problem, Tibshirani et al. (16) developed sparse nearest shrunken centroids (NSC) algorithm. NSC selects the minimal subset of genes by shrinking the class means to overall mean using lasso method, then classifies the data with unshrunken genes using DLDA method. Sparsity property of this algorithm leads to obtain simple, interpretable and lower variance classification model.

Microarray based algorithms are not directly applicable to RNA-Seq data since the discrete nature of RNA-Seq data is totally different than microarrays. Microarray data contain continuous data which are obtained from the log intensities of image spots, while RNA-Seq data contains discrete count data which represents the RNA abundances with the number of sequence reads mapped to a reference genome or transcriptome. During the past few years, much effort has been put into DE analysis of RNA-Seq data. Despite the importance of class discovery and class prediction analyses in gene expression data, there are still less advancements for RNA-Seq data until recently (17).

Preliminary studies applied logarithmic transformation to RNA-Seq counts and applied microarray based methods for DE (18–21). Even log transformation makes the data similar to microarrays; it gives more weights to highly expressed genes and less weight to weakly expressed genes. Following studies applied Poisson distribution based models to deal with the RNA-Seq count data (22, 23). However, Poisson distribution has a single parameter for both mean and variance. Nagalakshmi et al. (24) reported that Poisson modelling is only appropriate for RNA-Seq data, when the replicates are technical (i.e repeated measurements from same subjects). To make an inference, researchers have widely worked with biological replicates (i.e measurements taken from different subjects), where the variance of a gene exceeds its mean. Using biological replicates have arisen another problem; heteroskedasticity. Thus, Poisson models are inappropriate while working with heteroskedastic RNA-Seq data, and more care should be taken while modelling this data. Later studies applied negative binomial (NB) distribution due to its extra parameter to model overdispersion. Anders and Huber (25) proposed DESeq2 algorithm and later on extended it with DESeq2 (26). Robinson et al. (27) presented edgeR algorithm. Both algorithms use negative-binomial distribution in modelling. DESeq (also DESeq2) uses local regression, while edgeR uses a single constant value in modelling mean and variance relationship. Finally, Law et al. (28) proposed a different strategy and estimated the mean and variance relationship at observational level. The authors named this method as voom (variance modelling at observational level) and unlocked the use of mi-

croarray based methods for DE analysis. Voom transforms the RNA-Seq data by log counts per million (log-cpm) transformation and also provides observational weights for downstream analysis. Integration of voom method with limma algorithm provides the highest power, lowest false discovery rate and best controlling of type-I error as compared to other DE analysis methods (28).

In clustering and classification analyses of RNA-Seq data, two strategies are available similarly with DE analysis: (i) proposing novel algorithms based on discrete distributions such as NB, (ii) transforming data to make it distributionally closer to microarrays and apply microarray based algorithms (17). In classification part, Witten (8) presented Poisson linear discriminant analysis (PLDA) which is an extension of ordinary discriminant classifiers and NSC approach to the analysis of high-dimensional count data. Despite the advantages of PLDA such as being sparse and ability to discriminate counts, this algorithm is not able to deal with the overdispersion problem. To overcome this problem, Witten (8) suggested applying a power transformation for stabilizing the mean and variance relationship. Zararsiz et al. (29) used a variance stabilizing transformation (vst) and applied machine learning algorithms in classification of RNA-Seq data. The authors presented the application of these algorithms in MLSeq R/BIOCONDUCTOR package (30). In another study, Zararsız (17) presented voomDDA algorithms, which involves sparse voomNSC and non-sparse voomDLDA and voomDQDA algorithms. These algorithms integrate the voom method with powerful microarray classifiers including NSC, DLDA and DQDA.

Dong et al. (31) presented negative binomial discriminant analysis (NBLDA) classifier to extend PLDA and overcome the overdispersion problem. This algorithm uses a shrinkage method in predicting the extra overdispersion parameter. The major limitation of this classifier is that it is not sparse. Although, NBLDA overcomes the overdispersion problem, this algorithm involves the entire genes into the classification model and provides very complex models.

In clustering analysis, Witten (8) presented Poisson dissimilarity index as a distance measure for clustering RNA-Seq data. The authors applied this method with several normalization approaches, including total count, quantile and DESeq median ratio. Si et al. (32) presented model based clustering approach with expectation and maximization, and hybrid hierarchical clustering algorithms. Love et al. (26) studied the clustering of RNA-Seq data after vst and regularized log (rlog) transformations. Liu and Si (33) demonstrated the use of K-means, hierarchical, model-based and hybrid-hierarchical clustering algorithms on a real RNA-Seq data. The authors also applied a negative binomial mixture model in

model-based clustering method. Reeb et al. (34) investigated the effect of several dissimilarity measures for sample-based hierarchical clustering of RNA-Seq data. The authors provided plasmodes to guide researchers on selecting appropriate dissimilarity measure.

1.1. Contribution

In this thesis, we presented novel approaches for classification and clustering of RNA-Seq data. In classification part, we extended the work of Dong et al. (31) and focused on the sparse NBLDA classifier. We also provided sparse solutions for heteroskedastic RNA-Seq data. In this way, we overcome the overdispersion problem of PLDA and complexity problem of NBLDA. Both discrete and transformation based algorithms were provided for the clustering part. Here, we focused on NB distribution based and voom based dissimilarity indices and clustering algorithms. In brief, the major objectives of this thesis were two-fold:

- to present a sparse NBLDA classifier that will identify the best minimal subset of genes and predict the class conditions,
- to present novel clustering procedures based on discrete distributions and transformations.

We also demonstrated the applicability of the developed approaches in our self-developed MLSeq package.

1.2. Organization of This Thesis

The rest of the thesis is organized as follows:

'General Information': In this section, RNA-Seq is detailed in both technological and methodological view. The experimental pipeline of this technique and the generation of raw data are addressed here. Moreover, the preprocessing of this data and the currently used algorithms are explained in this section.

'Material and Methods': The proposed methodologies are given in details in this section. The performance assessment of the presented algorithms is mentioned. A comprehensive simulation study is designed and detailed in this section. Furthermore, real datasets are also used. The properties of these datasets and analysis workflow are addressed.

'Results': The results of both simulation and real datasets are revealed here.

'Discussion': Obtained results are discussed with other studies.

'Conclusion': The importance of this thesis is concluded with the obtained results. Final remarks are also given in this section.

2. GENERAL INFORMATION

Biological functions of a living organism are controlled by proteins which are produced using the information stored in genes as DNA. This genetic information is first transcribed into RNA and finally translated into proteins. The transcription of specific genes into collection of RNA molecules (called as transcriptome) specifies different cell types and proteins produced within cells, and cellular activities are regulated by produced proteins. Hence, the amount of RNA molecules and proteins are controlled by the expression level of related gene and environmental factors. These RNA molecules and expression level of genes are essential to understand cellular activities, development of cells and diseases (35).

The first gene expression studies used microarray technologies. However, RNA-sequencing became popular and frequently preferred for gene-expression studies in recent years with the advancement of next-generation sequencing (NGS) technology. A number of commercial NGS platforms are available which can be categorized into three main categories (17): (i) second generation platforms such as Illumina HiSeq, Roche 454 and ABI/SOLID, (ii) third generation platforms such as Ion Torrent, Pasific Bio and Complete Genomics, and (iii) fourth generation platforms such as Oxford Nanopore. Each platform has advantages and disadvantages in terms of computation time, accuracy, sequencing depth, output data size, etc. Furthermore, selected sequencing platform can be an important criteria for defining methods for downstream analysis and interpretation of the results. For this reason, appropriate sequencing platform should be defined by considering experimental goals, advantages and disadvantages simultaneously.

2.1. RNA Sequencing

RNA sequencing is used to obtain gene expression data from transcriptome by using one of NGS platforms. Recently, Illumina HiSeq platform became popular and standard sequencing platform for RNA sequencing studies. Although there might be differences in RNA sequencing workflow depending on sequencing platform, the main idea in the background is still same. Figure 2.1 shows a simple RNA sequencing workflow through Illumina platform where mRNA molecules are sequenced to reference genome. However, RNA sequencing has more detailed workflow starting from experimental design to feature counting. In this section, we detailed RNA sequencing workflow.

In the first step of RNA sequencing workflow, correct **experimental design** should be determined to obtain unbiased and accurate results. A good

experimental design considers basic principals; randomization, replication and blocking. There are two different types of replication that are *biological* and *technical* replications in an RNA sequencing. Technical replicate corresponds to two repeated measurements taken from the same subject. Biological replicate, on the other hand, corresponds to two measurements taken from different samples. Although both types of replicates can be preferred to increase statistical power, technical replicates generally preferred to increase sequencing depth and better detect differentially expressed genes. Liu et al. (36) showed that adding more biological replicates is better for increasing statistical power regardless of sequencing depth. Randomization and blocking are applied during sample preparation for randomly allocating samples to each block and groups.

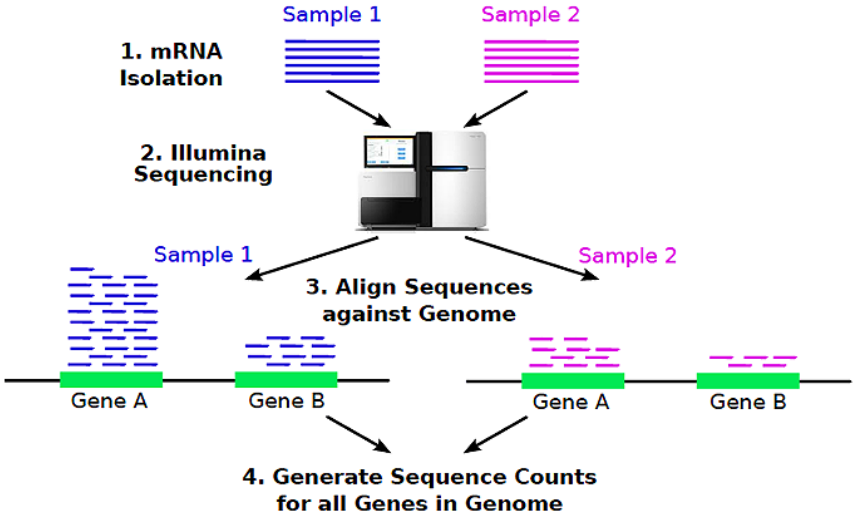


Figure 2.1. A basic RNA sequencing workflow through Illumina platform

Transcriptome is quite heterogeneous including different types of RNA molecules such as protein-coding mRNA and non-coding RNA species (e.g micro RNA, long non-coding RNA, ribozomal RNA, small RNA, etc.). Hence, it is important to define appropriate library preparation protocol for measuring RNA molecules of interest. For example, ribozomal RNA molecules accounts for 95% of the complete transcriptome within cell. If these molecules are not removed before library construction (i.e region of interest to where sequenced samples will be mapped) an important part of mapped reads will be associated with ribozomal RNA. Hence, less-abundant RNAs will not be able to correctly detected since sequencing reads of these molecules are relatively small. For this reason, in **sample and library preparation** step (Step 1 in Figure 2.1), unnecessary

RNA molecules are excluded and interested gene regions[1] are determined.

In **quality assessment and alignment** step (Steps 2 and 3 in Figure 2.1) isolated RNA molecules are sequenced using NGS platform, and millions of short sequence reads are obtained in a single run along with a quality score of each read. Next, raw sequencing reads are preprocessed, low quality reads are filtered, and filtered short reads are aligned to reference genome. Finaly, mapped reads are counted for each gene region in **feature counting** step (Step 4 in Figure 2.1).

In conclusion, a typical RNA sequencing experiment briefly consists of following steps:

1. RNA molecules are isolated from transcriptome and fragmented to an average length of 200 bases.
2. Fragmented RNA molecules are converted into complementary DNA (cDNA).
3. cDNA fragments are sequenced and aligned to reference genome.
4. Mapped read counts are obtained.

2.2. Data Structure

High-throughput technologies, e.g microarrays and RNA sequencing, return gene expression data matrix $\mathbf{X}$ where each row represents gene regions and each column represents samples. The main difference between microarray and RNA sequencing data is that microarrays returns gene expression data in continuous scale while RNA sequencing returns in discrete scale. Suppose mRNA molecules of three healthy and three diseased subjects are sequenced and aligned to 500 different reference gene regions. Table 2.1 shows an example of RNA sequencing data which might be obtained as detailed in Figure 2.1. Each cell in data matrix (x_{ij}) corresponds to total mapped read counts of j-th sample to i-th gene. These counts are related with the gene expression levels. Hence, the number of mapped read counts increases as the activity of related gene increases. However, mapped read counts depends on not only the gene expression level but also sequencing depth, gene length, quality of sequences, etc. Hence, raw counts can not be directly used as a measure of gene expression level unless counts are preprocessed for downstream analyses such as differential expression, classification, clustering etc. For example, raw counts should be normalized in order to remove between sample differences such as sequencing depth and compare each

[1] There is no need to define specific gene regions if whole exom sequencing is performed. In this case, sequenced reads mapped to any specific region of complete human DNA.

sample in the same scale. Similarly, raw counts might be transformed to be used in clustering and classification tasks which are proposed for continuous data.

Table 2.1. An example of RNA sequencing data.

Features	Healthy			Diseased			Total
	H1	H2	H3	D1	D2	D3	
Gene1	304	3	14678	92	177	1	97865
Gene2	0	0	438500	7600	47387	62	9472601
Gene3	14	26	972	48	714	184	74318
...	...	...	...	...	...	...	...
Gene500	0	1	72	14390	28400	187	162840
Total	6792	1376	12765432	151100	338405	9320	89366412

Over the last two decades, lots of methods have been developed for measuring gene expression levels from sequencing data and to understand the relation between gene expression levels, cellular activities and diseases. Furthermore, many machine learning algorithms are proposed for clustering and classification of RNA sequencing data. We will cover methods which are specifically proposed for differential expression, preprocessing and machine learning in the following sections.

2.3. Normalization

The total number of mapped reads might be very different for each sample and gene depending on the sequencing depth and gene length. For example, the mapped read counts to same feature of two different subjects might greatly differs depending on the sequencing depth. Similarly, the mapped read counts to different genes of the same subject might also differs depending on the gene length. As a result, two quantities $x_{i.} = \sum_{j=1}^{n} x_{ij}$ and $x_{.j} = \sum_{i=1}^{p} x_{ij}$ depend on experimental design in an RNA-Seq study, and this will yield technical biases in downstream analysis. Therefore, raw read counts should be normalized before continuing further analyses, e.g differential expression analysis for detecting significant genes, classification and/or clustering of samples via machine learning algorithms.

Previous microarray studies have repeatedly proven that normalization is a crucial step in differential expression analysis in order to decrease false positive results and obtain accurate estimates (37). Although RNA-Seq produces less noisy data comparing to microarrays, normalization is still an important issue

to be considered since it enables researchers to compare gene expression levels of two different samples which may have different sequencing depths. We call total number of reads of a sample, $x_{\cdot j}$, as *sequencing depth* or *library size*. If i-th gene is not differentially expressed, we expect that the ratio of expected counts $E(X_{ij})/E(X_{ij'})$ for subjects j and j' should be equal to the ratio of size factors $s_j/s_{j'}$ of the same subjects. Hence, it is obvious that one should estimate *size factors* in order to normalize raw counts. All normalization methods discussed in this thesis use a single size factor estimation s_j for each subject rather than using estimates s_{ij} for each cell, and mapped read counts for each gene is globally normalized using size factor s_j of that subject. Although it is possible to estimate different size factors for j-th sample on each feature, we generally assume size factor estimate of each sample is constant for all features in order to decrease the model complexity.

Data generation pipeline for RNA-Seq experiment is different from that of microarrays. For this reason, the normalization procedures –also called as *size factor estimation*– proposed for microarrays are not directly applicable to RNA-Seq data. Recent studies proposed several normalization techniques for RNA-Seq data. Among those, we will discuss and use *total count*, *median ratio* and *upper quartile* normalization methods. Other normalization methods such as trimmed mean of M-values, quantile, reads per kilobase per million mapped reads (RPKM), CuffDiff, PoissonSeq etc. are available through references (18, 22, 37–43).

Most of the proposed algorithms for genomics data are used to detect differentially expressed genes. However, further steps required for classification and clustering tasks while estimating size factors of training and test samples. The size factor estimation procedures for a test sample $\mathbf{x}^* = \{x_1^*, x_2^*, \ldots, x_p^*\}$ is not straightforward since we should estimate size factor s^* of a test sample by using training set parameters. Finally, normalized read counts are calculated by x_{ij}/s_j for training set samples and x^*/s^* for a test sample.

2.3.1. Total Count Normalization

A number of authors proposed normalizing read counts by scaling each sample to library sizes (6, 40, 41). One may estimate size factor of j-th sample as

$$s_j = \frac{\sum_{i=1}^{n} x_{ij}}{\sum_{i=1}^{p}\sum_{j=1}^{n} x_{ij}} = \frac{x_{\cdot j}}{x_{\cdot\cdot}} \tag{2.1}$$

Under independence assumption of samples and genes, total count estimate is derived from maximum likelihood estimate of cell means $\hat{\mu}_{ij} = x_{i.}x_{.j}/x_{..}$ as stated by Witten (8). Therefore, total count estimation is also called as *maximum likelihood estimate* in differential expression analysis. Size factor of a test sample $\mathbf{x}^*$ is estimated by $s^* = \sum_{i=1}^{p} x_i^*/x_{..}$ where $x_{..}$ is the grand total mapped reads of training set.

Total count estimation is frequently used in early RNA-Seq experiments; however, recent studies showed that it is not a very good estimate for size factors since a few outliers in j-th sample greatly inflates the estimated values, and overestimated size factors fails to detect differentially expressed genes (25, 26, 37, 44). For this reason, more robust methods such as *upper quartile* and *median ratio* were proposed for size factor estimations.

2.3.2. Upper Quartile Normalization

Bullard et al. (44) proposed normalizing read counts of each sample by scaling to 75th percentile of the counts for that sample. This normalization is robust to outliers since it takes upper quartiles into account rather than total counts. The size factor for j-th sample is calculated by $s_j = q_j / \sum_{j=1}^{n} q_j$ where q_j is the upper quartile of the read counts for j-th sample. Size factor of a test sample $\mathbf{x}^*$ is estimated by $s^* = q^* / \sum_{j=1}^{n} q_j$ where q^* is the upper quartile of a test sample over p features and q_j is the upper quartile of j-th sample obtained from training set.

2.3.3. DESeq Median Ratio Normalization

Another robust alternative to total count normalization for size factor estimation is the median-of-ratios method proposed by Love et al. (26). The size factors are estimated by $s_j = m_j / \sum_{j=1}^{n} m_j$ where m_j is defined by

$$m_j = \underset{i:\, G_i \neq 0}{\text{median}} \left\{ \frac{x_{ij}}{G_i} \right\}, \qquad G_i = \left(\prod_{j'=1}^{n} x_{ij'} \right)^{1/n} \tag{2.2}$$

Here G_i is the geometric mean of read counts for i-th feature and m_j is calculated by using the features having nonzero geometric mean. The size factor of a test sample $\mathbf{x}^*$ is then calculated by

$$s^* = \frac{m^*}{\sum_{j=1}^{n} m_j}, \qquad m^* = \underset{i:\, G_i \neq 0}{\text{median}} \left\{ \frac{x_i^*}{G_i} \right\} \tag{2.3}$$

where m_j and G_i are calculated from training set by using the equation 2.2. The denominator G_i is a pseudo-reference sample to which each sample is compared (45).

2.4. Transformation

Most of the proposed algorithms for differential expression analysis of RNA-Seq data are based on discrete distributions such as Poisson and negative binomial, and raw counts are directly used without transformation (26, 27, 46). However for other purposes such as clustering, classification and graphical representations, it might be a proper choice to work with transformed data due to several reasons. First, the variance of read counts for a gene is highly dependent on the expected count for that gene, and this results in the problem of heteroskedasticity for RNA-Seq data. This problem yields inaccurate results from clustering and classification tasks since a major part of these algorithms works well in normally distributed homoskedastic data. Secondly, the mapped read counts are heavily right skewed and overdispersed. Finally, discrete distributions are less tractable in terms of mathematical theory than that of normal distribution. As a result, the performance and the usefulness of discrete distributions in RNA-Seq studies tends to be limited (28).

The purpose of transforming counts is to obtain approximately normal and homoskedastic data. Therefore, transformation enables researchers to use statistical methods (e.g hierarchical clustering, linear discriminant analysis, heatmaps etc.), which are specifically developed for normally distributed microarray data, for RNA-Seq studies. The simplest transformation is possibly a *logarithmic* transformation $X \to Z$ by

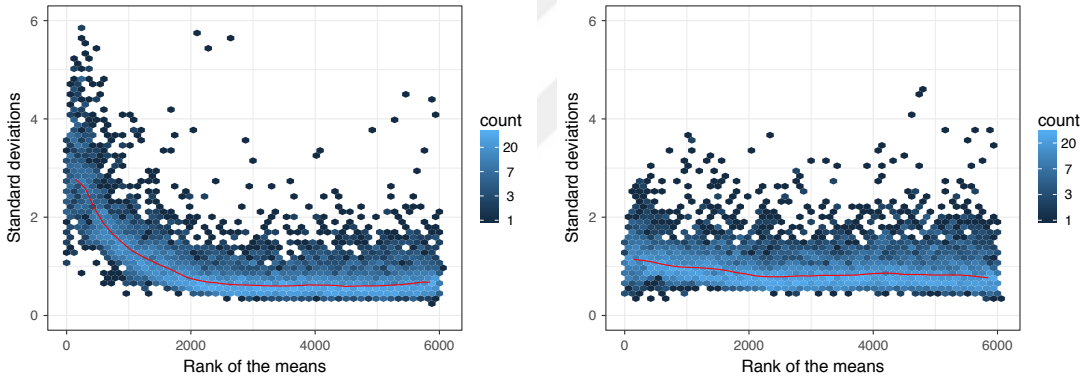
$$z_{ij} = \log_2 (x_{ij} + c), \qquad Z_{ij} \sim \text{Normal}\left(\mu_{ij}, \sigma_{ij}^2\right) \tag{2.4}$$

Here c is a small positive constant protecting the transformation from taking logarithm of zero. Furthermore, c plays a role for that distribution is shifted to the right. For this reason, the transformation 2.4 is also called as *shifted logarithmic* transformation. The logarithmic transformation brings right-skewed distribution to a approximately symmetric distribution. An important drawback of this transformation is that very high read counts have smaller weights while very small ones have undue weights in transformed space. Hence, it is more likely to observe increased false positive rates for detecting differentially expressed genes in logarithmic scale (28).

Recently, more robust and sophisticated transformation techniques have been proposed for RNA-Seq data. In this section, we will cover and discuss most common ones such as variance stabilizing (vst), regularized logarithmic (rlog), power and voom transformations.

2.4.1. Variance Stabilizing Transformation

The variance stabilizing transformation is one of the proposed algorithms for making the estimated variances independent of the means. When there is high overdispersion in data, the variance of $\log_2$ scaled counts is generally expected to be higher than its mean and this dependency decreases as mean increases. As can be seen from Figure 2.2, variance stabilizing transformation removes a major part of mean-variance dependency and transformed values have almost constant mean-variance trend. This figure is generated by using top $6,000$ genes of *lung cancer* data (see section 3.9 for details) after genes are sorted by their variances in descending order.



(a) **(b)**

Figure 2.2. The effect of variance stabilizing transformation on mean-variance trend – **(a)** Normalized counts by $\log_2(x+1)$ and **(b)** variance stabilizing transformed values

Consider a random variable X for mapped read counts with mean-variance relation $\sigma^2 = \mu + \phi\mu^2$ where ϕ is an overdispersion parameter. Anders and Huber (25) used the following parametrization for the relation between mean and overdispersion parameter

$$\phi = \phi_0 + \eta/\mu \tag{2.5}$$

where ϕ_0 is asymptotic overdispersion and η is an *extra-Poisson* parameter. The

variance as a function of mean is obtained by

$$\nu(\mu) = \sigma^2 = \mu + \mu^2\phi_0 + \mu\eta \tag{2.6}$$

Therefore, a variance stabilizing transformation $\tau(.)$ is a transformation yielding transformed value z such that

$$\tau(x) = z = \int^x \frac{1}{\sqrt{\nu(\mu)}}\,\mathrm{d}\mu \tag{2.7}$$

It is also possible to normalize raw counts within VST by replacing raw counts x with normalized counts $r = x/s$ and obtain transformed values $\tau(r)$. Here s is a size factor estimated separately for each sample. The transformed values z is now homoskedastic and can be used as an input to clustering and classification algorithms.

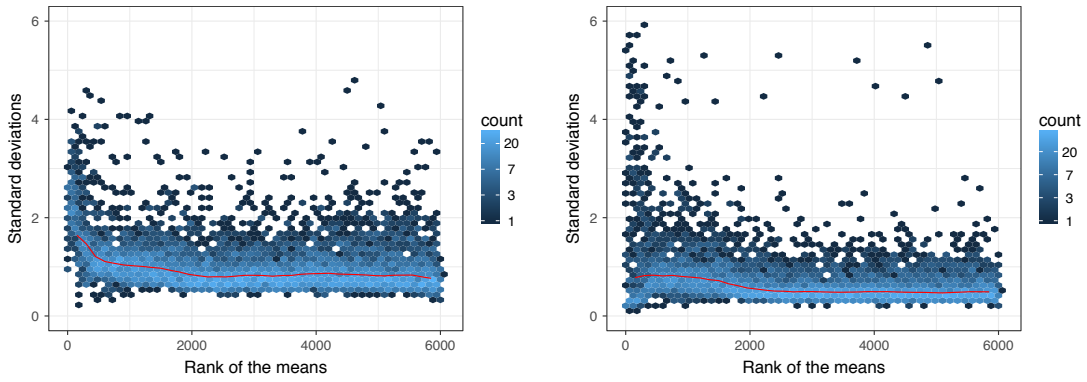
2.4.2. Regularized Logarithmic Transformation

The rlog transformation also transforms count data into $\log_2$ scale by using the same mean-variance trend as in vst. Transformed values are extracted from a generalized linear model (GLM) which is separately fitted for each gene by

$$\log_2 q_{ij} = z_{ij} = \beta_{i0} + \beta_{ij} \tag{2.8}$$

where β_{i0} is an intercept representing the expression level of i-th gene for all samples and β_{ij} is sample-specific expression level of genes. Here the term β_{ij} is shrunken towards 0 by using *ridge penalization* (or L2 penalization) while β_{i0} remained constant. The term q_{ij} is proportional to expected read counts and obtained by μ_{ij}/s_j where s_j stands for size factor of j-th sample (See Love et al. (26) for a complete steps of rlog transformation). Although rlog and vst serve for the same purpose, rlog performs better when size factors vary widely. Love et al. (26) suggested using rlog transformation when size factor estimates exceed 4.

An important limitation of rlog transformation is that it is computationally intensive comparing to vst since it fits a GLM for each gene. We generated Figure 2.3 for comparing the effect of rlog and vst on mean-variance trend. A subset of all samples is obtained by randomly selecting 10% of subjects from each class (i.e 58 and 55 subjects respectively) in order to decrease computation time. It can be seen from figure that rlog transformation is slightly better than vst for stabilizing the variance of genes which have lower mean. Furthermore, rlog transformation requires about 80 times longer computation time than vst for this



(a) **(b)**

Figure 2.3. Comparison of the effect of vst and rlog transfomation on mean-variance trend – **(a)** vst and **(b)** rlog transformation

example. Hence, the advantage of rlog transformation over vst might be ignored due to its intensive computation cost when sample size is large.

2.4.3. Power Transformation

Witten (8) proposed Poisson model for RNA-Seq classification. This model assumes that the mean and variance of read counts are equal. However, this assumption is violated since the variance of read counts is often greater than its mean. For this reason, a power transformation is applied to count data in order to bring variance around mean. Power transformation performs well when there is slight to moderate dispersion in data. Unlike vst and rlog transformation, power transformation does not take size factors into account. Hence, transformed values should be normalized before continuing with the classification. We will cover power transformation in details in section 3.1.

2.4.4. Voom Transformation

Variance modelling at the observational level (voom) aims to transform count data into continuous scale while obtaining precision weights along with transformed values. Law et al. (28) proposed voom transformation and suggested the use of statistical methods which are based on continuous data rather than working with discrete distributions. Their approach also focuses on mean-variance relationship of each gene similar to vst and rlog transformation. They have used transformed values along with its precision weights in differential expression analysis pipeline similar to microarrays and obtained desirable results for RNA-Seq data. The transformed values and precision weights are used as inputs to classification and clustering tasks. We give mathematical background of voom

transformation in section 3.4.

2.5. Classification of RNA-Seq Data

Gene expression data (also known as *transriptomic data*) can be used for several purposes such as inference and prediction. Classification, for example, is a popular machine learning technique which is used to predict class label of a new sample whose class label is unknown. There are basically two main steps of a classification problem; *learning* and *class prediction*. In learning step, the classification model learns patterns between gene expression levels and class labels (i.e between inputs and outputs) by using a predefined *training set*. Training set consists of gene expression – class label pairs. Here, we take gene expression levels as input and class labels as output to classification model. Finally, a function that maps input to output with the highest accuracy is obtained. In the second step, class labels of test samples are predicted on the basis of trained classification model. This type of machine learning technique is also called as *supervised learning* since trained model is optimized by using already known class labels (e.g diseased versus healthy, disease sub-types such as ALL and AML, tumor grades such as Grade 1, Grade 2 and Grade 3, etc.).

There are lots of classification algorithms in the literature where each performs well under specific conditions such as underlying probability distribution (discrete or continuous), data structure and dimensionality (i.e low or high dimensional in terms of number of samples and/or features). As we already know, RNA-Seq data is high-dimensional that is the number of features is much larger than the number of samples, $p \gg n$. We also know that most classification algorithms such as linear discriminant analysis and logistic regression will not work for such classification problems in high-dimensional space. As the number of features increases, the model becomes more complex, and this leads to difficulties in interpretation of the model as well as the *overfitting* problem. Although the fitted model perfectly classifies samples in training set, it will poorly classify an independent set of test samples which are not used while training the model. Hence, it is not a proper strategy to work with standard approaches for the high-dimensional setting even they are applicable. For this reason, a well-suited model for high-dimensional data aims to decrease model complexity via a *regularization parameter* or performs a *dimension reduction* while increasing the prediction accuracy of a test set (47, pp: 219–221).

We generally assume that only a small subset of all features is related with the class labels (or response) in the context of gene expression data. Although

more features than a selected subset may be associated with response in terms of biological rationale, it is suitable to put such an assumption to work with a simpler model for practical reasons. Some classification approaches such as Poisson linear discriminant analysis and nearest shrunken centroids have built-in variable selection criteria for selecting the best subset of all features. These models are called as *sparse* classifiers. The amount of sparsity or number of features included in the model is controlled by a regularization parameter (also named as *tuning* parameter), and the optimal value of tuning parameter is selected from a parameter space via cross-validation. Finally, trained model consist of features which contribute discrimination function the most. Same analogy can be used for dimension reduction in order to determine the optimal number of dimensions. Number of principal components for principal component classification (PCC) can be determined by considering the number of principal components as a tuning parameter. Hence, the optimal number of dimension is selected at a point which gives the highest classification accuracy.

In conclusion, high-dimensionalty is a major problem in RNA-Seq classification which limits the number of classification algorithms applicable. Standard approaches should be extended to high-dimensional settings and read counts should be preprocessed before classification. Besides, it is not possible to find a unique model which performs the best for RNA-Seq classification under all conditions. For this reason, we selected several classification algorithms which are based on discrete and continuous probability distribution, and compared accuracy of selected models. The mathematical background of selected classifiers will be discussed in section 3 in details.

2.5.1. Discrete Classifiers

RNA-Seq data can be directly modeled with discrete distributions such as Poisson and negative binomial. Recently, two popular classification algorithms are proposed for RNA-Seq classification. First, Witten (8) proposed a Poisson model (PLDA) for both classification and clustering of RNA-Seq data. Next, Dong et al. (31) extended classification part of Poisson model to negative binomial (NBLDA) case by considering an extra dispersion parameter. The proposed algorithms used discrete probability distributions in place of normal distribution within linear discriminant analysis. Hence, we may say that PLDA and NBLDA are extensions of linear discriminant analysis (LDA) to Poisson and negative binomial case. Although these algorithms are based on discrete distributions, raw read counts from RNA-Seq experiments should not be directly used within PLDA

and/or NBLDA. Raw counts are normalized by using one of normalization methods described in section 2.3 and normalized counts are used for classification purpose.

2.5.2. Continuous Classifiers

Continuous classifiers was popular and frequently used for classifying microarray based gene expression data. Their use in RNA-Seq data, on the other hand, may lead to inaccurate results without a proper transformation on read counts. Hence, transformed count data should be used within continuous classifiers. In section 2.4, we discussed transformation techniques which are proposed for RNA-Seq data. These methods are used to bring RNA-Seq data hierarchically closer to microarray data and enable the use of continuous classifiers which are specifically developed for microarray studies. There are dozens of continuous classifiers in the literature, however, we will cover nearest shrunken centroids (NSC) and its extension to voom transformation (voomNSC) in this thesis (16, 48, 49).

2.6. Clustering of RNA-Seq Data

Clustering analysis is a widely used machine learning task to explore hidden patterns in a multivariate data, and is simply a process of "gathering similar objects" together. Unlike classification, class labels are not known *a priori* and should be discovered from gene expression data. Because of that the class labels are unknown, clustering analysis is called as *unsupervised learning*. In a gene expression study, clustering analysis can be performed either on samples or features. Sample-based clustering, for example, can be performed in order to discover sub-types of a disease according to gene expression levels (2, 50) or different cell types in a tissue using single cell RNA-Seq (51, 52). Gene-based clustering, on the other hand, can be used to identify co-expressed features which are believed to be functionally related and included in the same biological pathway. By studying the genes in a pathway, biological causes of changes in gene expression levels can be explored. Furthermore, it allows to discover cellular processes and diseases that are related with the genes in a pathway. However, gene-based clustering leads to another limitation of clustering analysis. Most of RNA-Seq experiments include hundreds or thousands of features. When all of the features are included in clustering analysis, the model complexity dramatically increases while clustering performance decreases. For this reason, it is recommended to work with features that are specified as differentially expressed when gene-based clustering is considered.

Clustering of RNA-Seq data is not straightforward due to distributional and dimensional limitations as in classification. Therefore, we may perform cluster analysis either on transformed counts or use clustering approaches which are based on discrete probability distributions. For both gene-based and sample-based clustering analysis, samples and/or features are clustered into pre-defined number of clusters by using *similarity* or *dissimilarity* measures. There are many types of similarity and dissimilarity measures; however, we use one of widely used dissimilarity measure, that is Euclidean distance. Suppose $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ are p-dimensional vectors for j-th and j'-th subjects. The Euclidean distance between two subjects is

$$\Delta_{Euc}(\mathbf{x}_j, \mathbf{x}_{j'}) = \|\mathbf{x}_j - \mathbf{x}_{j'}\|^2 = \sqrt{\sum_{i=1}^{p} (x_{ij} - x_{ij'})^2} \tag{2.9}$$

Euclidean distance is the geometric distance between two points in p-dimensional space. This measure is highly sensitive to mean-shifting and scaling since squared difference of each feature is used in calculations. In gene expression studies, it is often suggested to use mean-centered data since the goal is to find similar patterns of gene expression profiles rather than the similar magnitudes of gene expression levels of each pair. Furthermore, gene expression levels can be centered and scaled before calculating dissimilarity measures (47).

Most clustering approaches perform well when data is approximately normally distributed. Hence, RNA-Seq data is transformed by using a proper transformation and clustering analysis is performed on transformed gene expression data. We used k-means and hierarchical clustering approaches for clustering transformed count data. Alternatively, read counts might be clustered by using clustering approaches which are based on discrete probability distributions. For instance, Poisson clustering is one of the approaches proposed for count data (8). This method calculates dissimilarity measures from Poisson model and applies hierarchical clustering by using *Poisson dissimilarity measures*. Poisson clustering is discussed in section 3.5. Negative binomial clustering is performed using dissimilarity measures obtained from edgeR package. edgeR performs multidimensional scaling to obtain distances between each pair of RNA samples which corresponds to leading log-fold-changes (53). Finally, the calculated dissimilarity matrix is used within hierarchical clustering similar to Poisson clustering routines.

2.6.1. K-means Clustering

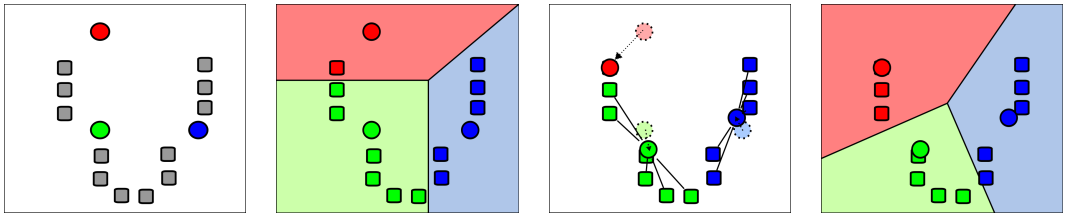
The K-means clustering aims to split samples/features into K subsets by minimizing the within-cluster sum of squares (i.e cluster variances). Suppose we perform sample-based clustering. Given a set of observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ where each element is a p-dimensional vector of gene expression levels, the algorithm determines K clusters by minimizing

$$\sum_{k=1}^{K} \sum_{j \in I_k} \|\mathbf{x}_j - \bar{\mathbf{x}}_k\|^2 \tag{2.10}$$

where $I_k \subset \{1, 2, \ldots, n\}$ is a subset of sample indices that belongs to class k and $\bar{\mathbf{x}}_k = n_k^{-1} \sum_{j \in C_k} \mathbf{x}_j$ is a p-dimensional mean vector of k-th cluster. The algorithm starts with an initial cluster means, which is generally taken as random points in p-dimensional space, and continues with the following steps until it converges:

1. Start with an initial cluster centroids (Figure 2.4a),
2. Assign each sample to one of K clusters whose cluster centroid is closest to that sample (Figure 2.4b),
3. Re-calculate cluster centroids after all samples are assigned one of K clusters (Figure 2.4c),
4. Re-assign each sample's cluster according to cluster centroids calculated at step 3 (Figure 2.4d),
5. Repeat steps 3 – 4 until no more change in cluster memberships take place.

Here, K is pre-defined number of clusters. The optimal number of clusters is another important criteria in K-means clustering, and it should be carefully determined. We assume that optimal value of K is already known and this value is provided in real data examples in this thesis. The graphical representation of K-means clustering algorithm is given in Figure 2.4.



(a) **(b)** **(c)** **(d)**

Figure 2.4. Steps of K-means clustering algorithm.

2.6.2. Hierarchical Clustering

Hierarchical clustering aims to build a hierarchy of clusters instead partitioning data into K subsets as in K-means clustering analysis. The hierarchy between samples/features can be constructed by using either agglomerative or divisive strategy. In agglomerative strategy, each sample is considered as a cluster, and samples are successively merged until all samples are collapsed into one big cluster (i.e complete data). Divisive strategy works in the opposite way such that all samples are considered as one big cluster, and samples are split into smaller clusters until each cluster consists of a single observation. The hierarchy of clusters can be displayed by using a dendrogram graph as in Figure 2.5. This figure is generated using cervical cancer data which includes gene expression levels of 29 tumor and 29 non-tumor samples from RNA-Seq experiment. More details about data set is given in section 3.9.

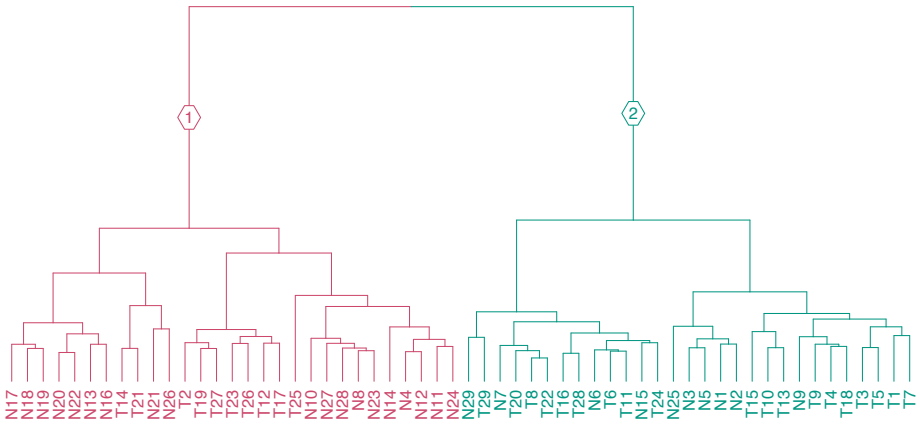


Figure 2.5. Hierarchical clustering dendrogram

Divisive strategy is not cost-effective, thus, it is not preferred in gene expression studies. We use agglomerative type of hierarchical clustering in this thesis. The merges in agglomerative strategy are determined using dissimilarity measures. When two clusters are required to be merged into one cluster, we need to calculate dissimilarities between clusters by using linkage methods. Some of well known linkage methods are illustrated in Figure 2.6. Selection of linkage methods may lead to very different clustering results. Single linkage considers dissimilarity measure between two nearest neighbours of each cluster while complete linkage considers two farthest neighbours. Centroid and average linkage methods, on the other hand, considers all samples in a cluster while calculating dissimilarities between two clusters. Centroid linkage uses cluster centroids and calculates the distance between centroids. Finally, average linkage uses dissimi-

larities of each pair from two clusters and takes its average as a dissimilarity of two clusters.

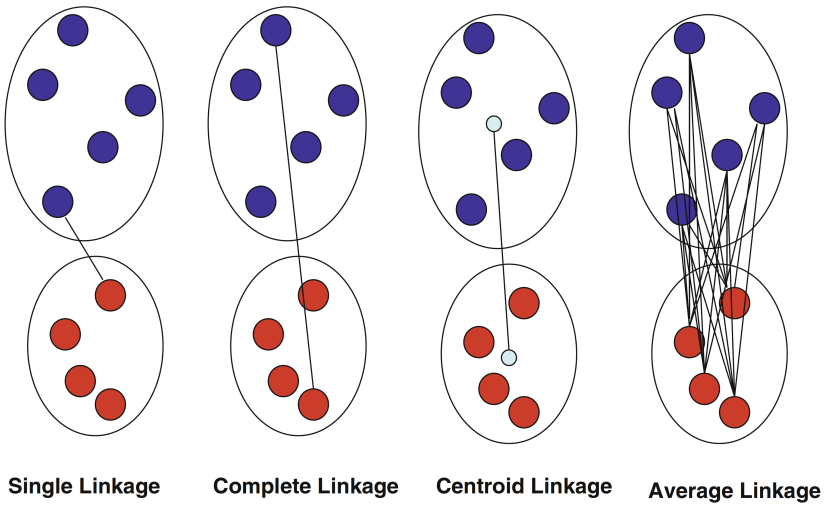


Figure 2.6. The linkage methods which are used to determine dissimilarities between clusters[2]

Hierarchical clustering can be performed using several dissimilarity measures. We used Euclidean distance within hierarchical clustering. As we already discussed, Euclidean distances are sensitive to non-normal data points and outliers. Hence, $\log_2(x+1)$ transformed values are used in clustering. Furthermore, we used all features in the model for graphical purposes; however, using only a subset of differentially expressed features may give better clustering performances. The clusters are obtained by cutting dendrogram at some level in order to get desired number of clusters. For instance, we cut tree in Figure 2.5 to find two clusters from cervical cancer data.

2.6.3. Visualizing RNA-Seq Data via Heatmaps

A heatmap is a graphical representation of 2-dimensional data where values are indicated by colors with changing intensities. Heatmaps can be used to make a quick look into data and understand complex patterns within data matrix. In gene expression studies, heatmap is a popular graphical representation of gene expression levels of subjects belonging to different clusters. A dendrogram is combined with a heatmap and gene expression levels are represented in different colors on a color palette in order to explore possible clusters. Figure 2.7 shows a heatmap of cervical data where gene expression levels are first normalized by

[2] This figure is copied from Datta and Nettleton (47), page: 200.

deseq median ratio and then transformed by $\log_2(r+1)$ where r is normalized counts and 1 is a pseudo-count[3].

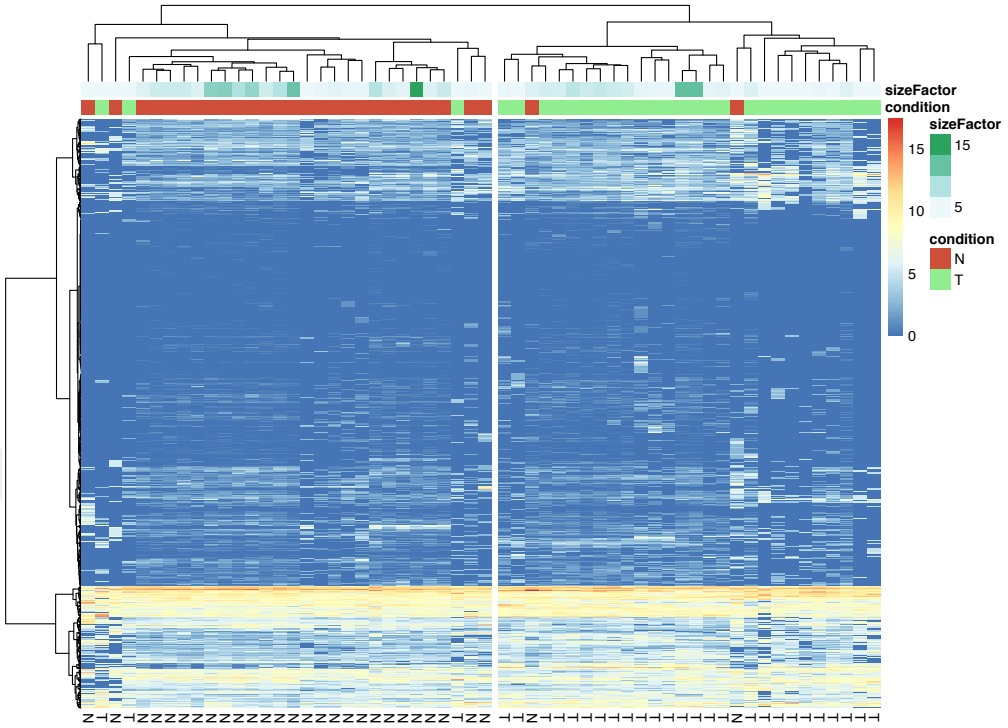


Figure 2.7. A heatmap of gene expression data of *cervical cancer* data. All features are included.

The resulting heatmap from gene expression data is affected by underlying distribution and number of features. Each value of a data matrix is colored proportional to its magnitude, hence, outliers and non-normal data points may lead to uninformative colors on a heatmap. We selected top 50 differentially expressed features and draw heatmap again as in Figure 2.8. When only a subset of differentially expressed features are used for clustering analysis, the relationship between samples and features can be seen more clearly.

Another use of heatmap is to discover similarities between samples by using a heatmap of sample-to-sample distances. Figure 2.9 shows distances between samples where 10 subjects are randomly selected from complete data and 50 differentially expressed features are selected based on selected samples.

[3] See function '`normTransform`' in R/Bioconductor package DESeq2 for more details.

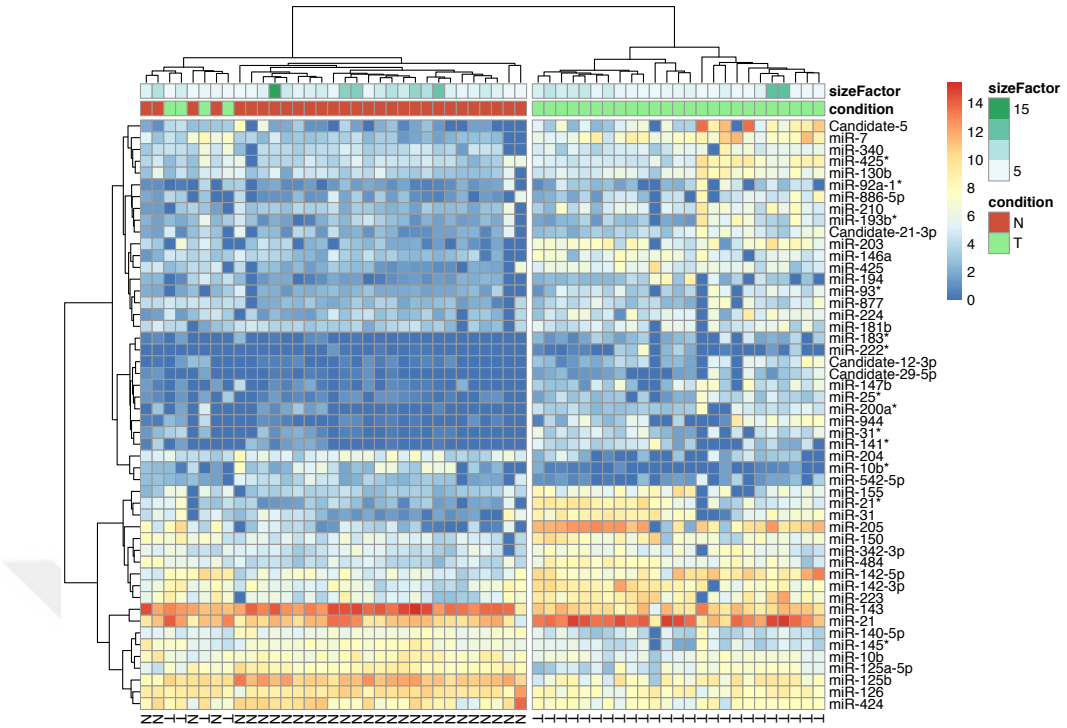


Figure 2.8. A heatmap of gene expression data of *cervical cancer* data. Top 50 differentially expressed features are included.

2.6.4. Assessing Cluster Performance and Within Cluster Consistency

Assessing clustering accuracies is not simple and straightforward as in classification because true class labels are not used while clustering samples and/or features. In clustering, Rand index (or Rand measure) is used to measure similarity between two clustering results (i.e predicted clusters from two different methods). Rand index is calculated from a contingency table of two clustering results. Its adjusted version, called adjusted Rand index, is also proposed for correcting chance within clustering. From a mathematical point of view, Rand index is related with accuracy measure in classification. We used adjusted Rand index for evaluating clustering accuracy. Since we know class labels, adjusted Rand index is calculated using similarities between predicted clusters and true class labels. Rand index takes value between 0 and 1 where 0 means there is no similarity between clustering results while 1 means clustering models perfectly agree. We skip mathematical background of Rand index, however, detailed background can be found in related papers (54, 55).

Beside clustering accuracy, it is important to measure within cluster consistency. A graphical approach, called silhouette, is proposed to measure cluster

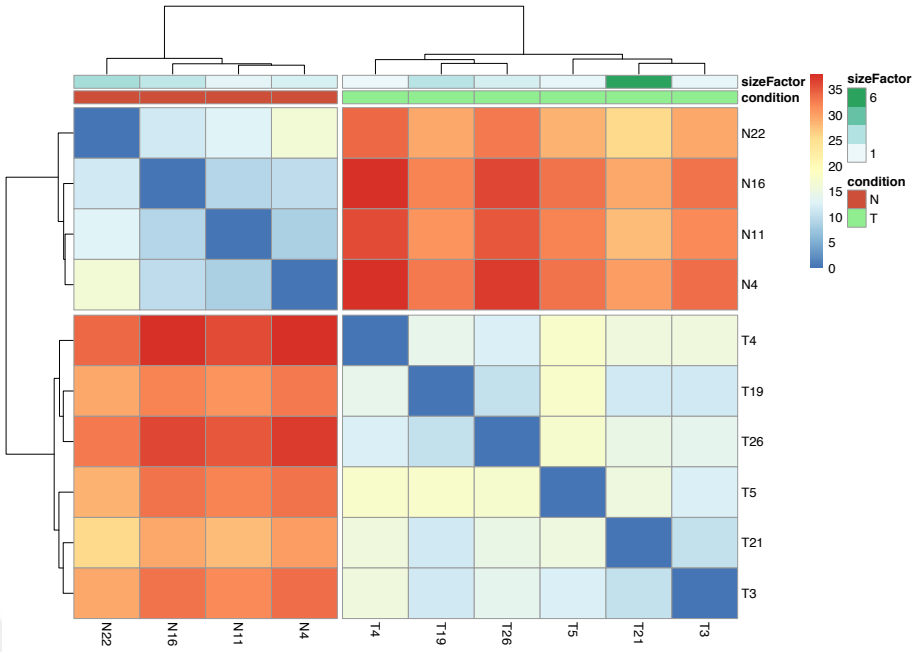


Figure 2.9. A heatmap of sample-to-sample distances of *cervical cancer* data. A subset of 10 samples is randomly selected and top 50 differentially expressed features are used through clustering analysis.

consistency (56). A silhouette measure s_i is calculated for each sample by

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2.11}$$

where a_i denotes the average distance between i-th sample and all other samples in its cluster, b_i denotes the smallest average distance between i-th sample and all the remaining samples from other clusters. Hence, s_i measures both how well i-th sample belongs to its cluster and other clusters. Equation 2.11 can be also written as

$$s_i = \begin{cases} 1 - a_i/b_i, & \text{if} \quad a_i < b_i \\ 0, & \text{if} \quad a_i = b_i \\ b_i/a_i - 1, & \text{if} \quad a_i > b_i \end{cases} \tag{2.12}$$

It is clear from equation 2.12 that s_i takes values within $[-1, 1]$. If $s_i = 1$, then i-th sample is well assigned to its cluster. If $s_i = -1$, then i-th sample is poorly assigned to its cluster and it should be assigned to one of remaining clusters. Finally, if $s_i = 0$, then i-th sample is located on the boundary of two clusters. An average silhoutte measure is obtained from all samples and it is used to measure overall within cluster consistency of clustering method.

3. MATERIAL and METHODS

A number of early RNA-Seq experiments analysed read counts using statistical methods which are specifically proposed for microarrays. Although RNA-seq counts are indicators for expression level of a specific gene, it was obviously not a good strategy to analyze count data with statistical methods which are based on continuous underlying distributions. As a result, the accuracies from fitted model was not satisfactory. Later studies addressed discrete distributions for the analysis of RNA-seq experiments.

Let $\mathbf{X}$ be a p-by-n matrix of mapped read counts from an RNA-Sequencing experiment, with n samples in the columns and p features in the rows as described in section 2.2. X_{ij} is the random variable indicating the total number of mapped read counts for gene i in observation j, and x_{ij} is the observed value for random variable. Since RNA-Seq experiments produce discrete count data, the mapped read counts for X_{ij} should follow a discrete probability distribution. We may fit count data to well-known discrete distributions such as Poisson and negative binomial. However, most of the statistical methods proposed for classification and/or clustering purposes assume the underlying distribution of data is normal. Hence, transforming raw counts using an appropriate transformation (e.g vst, rlog, log-cpm, voom, etc.) and modeling transformed counts with continuous distributions is still popular and hot topic in classification. For this reason, we will cover the mathematical background of both discrete (e.g Poisson linear discriminant analysis, Negative Binomial linear discriminant analysis) and continuous (e.g SVM, NSC, and voomNSC) classifiers in details throughout this chapter.

3.1. Poisson Linear Discriminant Analysis (PLDA)

One of the most popular discrete distribution which is used for modelling RNA-Seq data is the Poisson distribution (6, 22, 44, 57). Witten (8) proposed a Poisson log linear model for RNA sequencing data,

$$X_{ij} \sim \text{Poisson}\left(\mu_{ij}\right), \quad \mu_{ij} = s_j g_i \tag{3.1}$$

where s_j is the *size factor* of j-th sample and g_i is the *gene-length* of i-th feature. The proposed model 3.1 is able to account variability in raw counts using total number of reads for each sample via s_j term and total number of read for each feature via g_i term. In order to overcome identifiability problems, one should satisfy that $\sum_{j=1}^{n} s_j = 1$. An RNA-Seq experiment is often performed on samples taken from several subsets, say K *classes* or *disease subgroups*. We expect that

some of the genes are responsible for specific diseases, i.e the expression level of such genes in diseased subjects should be different than that of controls (or healthy subjects). Furthermore, we should observe that some features would have significantly different number of mapped reads as a result of gene expression levels. Hence, we may need to introduce class effects into 3.1. Some authors have extended 3.1 as in 3.2:

$$X_{ij} \mid y_j = k \sim \text{Poisson}\left(\mu_{ij} d_{ik}\right), \quad \mu_{ij} = s_j g_i \tag{3.2}$$

where y_j is the class of the j-th sample and $y_j = \{1, 2, \ldots, K\}$. The class specific parameter d_{ik} is called as *differential expression* or *offset* parameter which can be interpret how much the observed counts of i-th gene differs from expected (or baseline) counts for k-th class. The extended model 3.2 has two main parameters to be estimated, i.e Poisson mean μ_{ij} and offset parameter d_{ik}. The parametrization in 3.2 has a simple interpretation such that,

- estimate Poisson mean μ_{ij} by taking sample and gene-wise variations into account. This estimation is simply the baseline expected counts for random variable X_{ij},
- regulate baseline estimations using offset parameter d_{ik} for each class $k = 1, 2, \ldots, K$.

We first start estimating Poisson parameter μ_{ij}. We assume that features and samples are independent from each other. Under independence assumption, the Poisson mean for X_{ij} is estimated using maximum likelihood method (MLE) as $\hat{\mu}_{ij} = x_{i.}x_{.j}/x_{..}$ (58). Introducing s_j's and g_i's into estimated Poisson mean and combining this with the constraint $\sum_{j=1}^{n} \hat{s}_j = 1$ result in estimates of $\hat{s}_j = x_{.j}/x_{..}$ and $\hat{g}_i = x_{i.}$. Several studies have used MLE method for estimating size factors, however, it is obvious from the equation $x_{.j}/x_{..}$ that the estimated size factors are not robust to outliers since a few cells with very high counts may greatly inflates the size factor estimates (6, 8, 40). Furthermore, the inflated size factor estimates will lead to biased results for downstream analyses as pointed out by a number of authors (25, 37, 44). Because of this reason, several robust methods have been proposed for estimating size factors. These methods have been discussed in details in Section 2.3, hence, we continue by considering one of proposed methods.

The order of estimating Poisson log linear model parameters is an important issue that we should point out. Witten (8) follows an order that in the first step the Poisson mean is estimated using model 3.1; then, class differences have

been taken into account by estimating offset parameter d_{ik} from model 3.2. This analogy aims to account sample and feature-wise variability before measuring the class differences on the observed counts. The maximum likelihood estimates of $\hat{\mu}_{ij}$ is used to estimate offset parameter as,

$$\hat{d}_{ik} = \frac{\sum_{j \in I_k} x_{ij}}{\sum_{j \in I_k} \hat{\mu}_{ij}} \qquad i = 1, 2, \ldots, p \quad j = 1, 2, \ldots, n \tag{3.3}$$

where $I_k \subset \{1, 2, \ldots, n\}$ is a subset of sample indices that belongs to class k. The estimated offset parameters in 3.3 now has a simple interpretation such that

- if $\hat{d}_{ik} > 1$, then the i-th feature (gene, exom, etc.) is over-expressed in class k yielding that the number of mapped reads of corresponding feature is greater than the baseline in k-th class,
- if $\hat{d}_{ik} < 1$, then the i-th feature is under-expressed in class k yielding that the number of mapped reads of corresponding feature is less than the baseline in k-th class, and
- if $\hat{d}_{ik} = 1$, then the i-th feature is not differentially expressed in class k yielding that the number of mapped reads of corresponding feature is equal to the baseline in k-th class.

The offset parameter $\log(d_{ik})$ is normally distributed with mean 0 and a constant standard deviation σ_d for all classes, i.e $d_{ik} \sim \text{Normal}(0, \sigma_d^2)$. Finally, we fit model 3.2 by introducing estimated model parameters as below:

$$X_{ij} \mid y_j = k \sim \text{Poisson}\left(\hat{s}_j \hat{g}_i \hat{d}_{ik}\right), \quad \hat{\mu}_{ij} = \hat{s}_j \hat{g}_i \tag{3.4}$$

The fitted model 3.4 has a limitation when $\sum_{j \in I_k} x_{ij} = 0$. In this case, the estimated value of $\hat{d}_{ik}$ equals zero making downstream analysis indefinite for classification task as can be clearly seen in equation 3.9. This situation is likely to be occurred when the true mean for i-th feature is very small. We can overcome this limitation by putting a $\text{Gamma}(\beta, \beta)$ prior (a distribution having the same shape and scale parameters) on d_{ik} in the fitted model 3.4 (8). The posterior distribution for d_{ik} is then $\text{Gamma}\left(\sum_{j \in I_k} x_{ij} + \beta, \sum_{j \in I_k} \hat{\mu}_{ij} + \beta\right)$ and the posterior mean becomes

$$\hat{d}_{ik} = \frac{\sum_{j \in I_k} x_{ij} + \beta}{\sum_{j \in I_k} \hat{\mu}_{ij} + \beta}. \tag{3.5}$$

It is possible to put any arbitrary value for Gamma parameter β. We assume that $\beta = 1$ throughout this thesis.

3.1.1. Power Transformation on Count Data

An RNA-Seq experiment may consist of count data from either technical or biological replicates. Although many data sets of the early RNA-Seq studies include counts from technical replicates (i.e sequence reads from same source of RNA repeatedly taken from same individual) or biological replicates (sequence reads taken from different individuals). Previous RNA-Seq experiments revealed that variation in technical replicates are generally lower comparing to biological replicates (6, 27). The mapped read counts may be fitted by Poisson distribution assuming that the mean and variance of read counts are equal when technical replicates are taken. However, if biological replicates are taken, it is more likely to have more variation in read counts which causes the variance exceeds mean. For this reason, it may not be a proper choice using Poisson distribution since there is overdispersion in the data. Hence, we may either choose fitting data by Negative Binomial (NB) distribution (will be discussed in the next section) or apply power transformation on observed counts as proposed by Witten (8).

The power transformation is able to remove overdispersion when data is slightly or moderately overdispersed. Using maximum likelihood method and independence assumption, we obtain a transformed random variable $X_{ij}^{\dagger} \leftarrow X_{ij}^{\alpha}$ satisfying that

$$\sum_{i=1}^{p}\sum_{j=1}^{n} \frac{\left(x_{ij}^{\dagger} - x_{.j}^{\dagger}x_{i.}^{\dagger}/x_{..}^{\dagger}\right)^2}{x_{.j}^{\dagger}x_{i.}^{\dagger}/x_{..}^{\dagger}} \approx (n-1)(p-1) \tag{3.6}$$

where α is taken in the interval $(0, 1]$ and $(n-1)(p-1)$ is the *degrees of freedom*. This transformation is simply a goodnes-of-fit test for model 3.1 using the maximum likelihood estimate (or *total count* normalization method) of size factor $\hat{s}_j = x_{.j}^{\dagger}/x_{..}^{\dagger}$. The optimal value of α is selected by using a grid search within the interval $(0, 1]$.

The transformed counts in 3.6 is now able to be fitted by Poisson model. Although the transformed values are not integers and have decimal points, it still preserves the discrete nature since two consecutive integer counts yield again two consecutive real value (8, 57). Furthermore, there is no chance to get a transformed data point between these two consecutive values. The optimal value of α is determined using a grid search in the interval $(0, 1]$ and the value that

approximates 3.6 the best is chosen as the power transformation parameter.

3.1.2. Classifying New Samples

In a classification task, we want to predict which class a new sample might belong. Let $\mathbf{x}^* = \{x_1^*, x_2^*, \ldots, x_p^*\}$ be a test observation whose class to be predicted using the fitted model. Also let y^* be the true class of a test sample which is unknown. We may obtain posterior probability of a test sample being in class k by Bayes' rule,

$$P\left(y^* = k \mid \mathbf{x}^*\right) \propto f_k\left(\mathbf{x}^*\right) \pi_k \tag{3.7}$$

where π_k is the prior probability for k-th class. Here, the choice of density function $f_k(.)$ determines the type of discriminant analysis. For example, if we assume that data is normally distributed with a common covariance matrix and class-specific means, the classification method becomes standard linear discriminant analysis (LDA) (59). We assume that data follow a Poisson distribution and features are independent. Although some features are expected to be correlated, the independence assumption of features is frequently made when data is high dimensional (16, 48, 59), i.e the number of features are very large comparing to number of samples ($p \gg n$). In order to estimate posterior probabilities in 3.7, $f_k(\mathbf{x}^*)$ and π_k are need to be estimated. Poisson model 3.2 can be written for a test sample $\mathbf{x}^*$ such that

$$X_i \mid y^* = k \sim \text{Poisson}\left(\mu_i d_{ik}\right), \quad \mu_i = s^* g_i \tag{3.8}$$

where estimated values for g_i and d_{ik} are imported from fitted model 3.4 of the training set. The size factor of a test sample, s^*, is also estimated using required values from training data. See Section 2.3 for more details on how size factors are estimated for a test sample from training data.

The prior probabilities might be taken equal such that $\hat{\pi}_k = 1/K$. As an alternative, we could define each class having prior probability that is proportional to the number of individuals in k-th class in the training set as below:

$$\hat{\pi}_k = \frac{\#\text{ of samples in class k}}{n}$$

Now, we introduce the estimated parameters s^*, $\hat{g}_i$ and $\hat{d}_{ik}$ into 3.7 and get the

discriminant scores as,

$$\begin{aligned}\log P\left(y^* = k \mid \mathbf{x}^*\right) &= \log \hat{f}_k(\mathbf{x}^*) + \log \hat{\pi}_k + c \\ &= \sum_{i=1}^{p} x_i^* \log \hat{d}_{ik} - \hat{s}^* \sum_{i=1}^{p} \hat{g}_i \hat{d}_{ik} + \log \hat{\pi}_k + c' \end{aligned} \tag{3.9}$$

where c and c' are some constants having no contribution to classification. Equation 3.9 showed that the estimated scores are linear function of observed counts x_i, so that the method is called as *Poisson linear discriminant analysis* (8). Again, we see from the equation that observed counts for i-th feature make no contribution to discriminant scores if and only if $\hat{d}_{ik} = 1$ for all classes. Hence, we say that the i-th feature is not differentially expressed among classes. With this property, PLDA algorithm is able to select differentially expressed genes by filtering genes whose offset parameters equal to 1 for all classes.

3.1.3. Sparse Poisson Linear Discriminant Analysis

The sparse PLDA classifier removes insignificant genes from the model when the estimations from equation 3.3 are used. We use estimations from equation 3.5. These estimates yields $\hat{d}_{ik} \neq 1$ for all classes, hence, all features are included in the model even if the estimated differential expression parameters are very close to 1. Witten (8) proposed a shrinkage algorithm similar to nearest shrunken centroids algorithm (16, 48). This algorithm aims to shrink $\hat{d}_{ik}$ towards 1 by using a threshold ρ as below:

$$\hat{d}_{ik} = \begin{cases} \dfrac{a}{b} - \dfrac{\rho}{\sqrt{b}}, & \text{if} \quad \sqrt{b}\left(\dfrac{a}{b} - 1\right) > \rho \\ \dfrac{a}{b} + \dfrac{\rho}{\sqrt{b}}, & \text{if} \quad \sqrt{b}\left(1 - \dfrac{a}{b}\right) > \rho \\ 1, & \text{if} \quad \sqrt{b}\left|1 - \dfrac{a}{b}\right| < \rho \end{cases} \tag{3.10}$$

where $a = \sum_{j \in I_k} x_{ij} + \beta$, $b = \sum_{j \in I_k} \hat{\mu}_{ij} + \beta$ and $\rho \geq 0$. As the threshold parameter decreases towards 0, less features are removed from fitted model.

The upper bound of threshold parameter is determined using the last line of equation 3.10 for each feature in each class $k = 1, 2, \ldots, K$ such that

$$\rho_{up} = \max_{ijk}\left(\sqrt{b}\left|1 - \frac{a}{b}\right|\right) \tag{3.11}$$

Finally, the optimal value for threshold parameter (ρ) which gives the highest

classification accuracy is determined using a grid search within $[0, \rho_{up}]$.

3.2. Negative Binomial Linear Discriminant Analysis (NBLDA)

Negative binomial (NB) distribution is a generalization of Poisson distribution for modeling count data when variance is larger than mean, i.e there is overdispersion in the data. Let X_{ij} be, as in Poisson model 3.2, a random variable of observed read counts for j-th sample in i-th feature. We consider the negative binomial model

$$X_{ij} \sim \mathrm{NB}\left(\mu_{ij}, \phi_i\right), \quad \mu_{ij} = s_j g_i d_{ik} \tag{3.12}$$

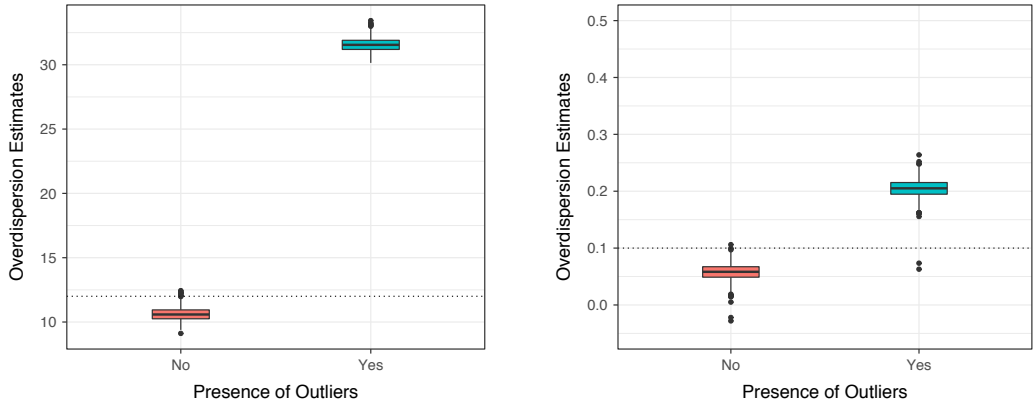
where ϕ_i is the overdispersion parameter for the i-th feature and the remaining parameters are the same as in Poisson model in terms of interpretation and estimation (31). The probability density function of X_{ij} is then,

$$P\left(X_{ij} \mid y_j = k\right) = \frac{\Gamma(x_{ij} + \phi_i^{-1})}{x_{ij}!\,\Gamma(\phi_i^{-1})} \left(\frac{\omega_{ijk}\phi_i}{1 + \omega_{ijk}\phi_i}\right)^{x_{ij}} \left(\frac{1}{1 + \omega_{ijk}\phi_i}\right)^{\phi_i^{-1}} \tag{3.13}$$

where $\omega_{ijk} = s_j g_i d_{ik}$. The mean and variance of random variable X_{ij} are obtained as μ_{ij} and $(\mu_{ij} + \mu_{ij}^2\phi_i)$ using method-of-moment estimators. As ϕ_i approaches 0, the negative binomial model 3.12 approximates to the Poisson model 3.2.

3.2.1. Power Transformation on Count Data

In section 3.1.1, we discussed the effect of power transformation on the prediction accuracy of the PLDA model. Power transformation is able to improve PLDA model accuracy when data is slightly or moderately overdispersed; however, its effect on highly (or extremely) overdispersed data is negligible. Even though NBLDA is an appropriate choice for modeling highly overdispersed data, it might perform poorly in presence of outliers (i.e very high read counts due to deeply sequenced samples) and/or extreme overdispersion, e.g $\phi > 5$. We simulated an p-by-n dimensional dataset where $n = 500$, $p = 20$ and each feature follows a negative binomial model 3.12 with mean $\mu = 20$ and overdispersion $\phi = 12$ for all X_{ij}s. Furthermore, we assume that size factors, s_j, are same across samples, each feature has approximately same gene length, g_i, and features are not differentially expressed among classes, i.e $d_{ik} = 1$ for all features. Finally, we replaced some of the values with outliers in order to see how gene-wise dispersions are over-estimated. The power transformation is performed as in PLDA using equation 3.6.



(a) **(b)**

Figure 3.1. The effect of outliers on gene-wise overispersion estimates.

Figure 3.1 shows gene-wise overdispersion estimates for untransformed (Figure 3.1a) and power transformed (Figure 3.1b) read counts in presence of outliers. Dotted horizontal lines represent true overdispersion in Figure 3.1a and moderate overdispersion threshold in Figure 3.1b. When there are outliers in the data, overdipersions are overestimated. Furthermore, it can be seen that power transformation is able to reduce overdispersion below moderate level when there is no outlier, however, overestimation problem remains when there are deeply sequenced samples in the data. To overcome this problem, either we may exclude outliers from classification or apply power transformation and fit mapped read counts to negative binomial model by taking moderate overdipersions into account. Hence, we may observe increased classification accuracy after power transformation for negative binomial model when there are deeply sequenced samples in the data.

3.2.2. Classifying New Samples

Classifying a test sample $\mathbf{x}^* = \{x_1^*, x_2^*, \ldots, x_p^*\}$ is almost the same with Poisson model except that an extra dispersion parameter, ϕ_i, should be estimated for negative binomial model. The discriminant score for negative binomial model is obtained using Bayes' rule (as in equation 3.7) as below:

$$\begin{aligned} \log P(y^* = k \mid \mathbf{x}^*) &= \log \hat{f}_k(\mathbf{x}^*) + \log \hat{\pi}_k + c \\ &= \sum_{i=1}^{p} x_i^* \left[\log \hat{d}_{ik} - \log\left(1 + \hat{\omega}_{i*k}\hat{\phi}_i\right) \right] \\ &\quad - \sum_{i=1}^{p} \hat{\phi}_i^{-1} \log\left(1 + \hat{\omega}_{i*k}\hat{\phi}_i\right) + \log \hat{\pi}_k + c' \end{aligned} \tag{3.14}$$

where $\hat{\omega}_{i*k} = \hat{s}^* \hat{g}_i \hat{d}_{ik}$. Furthermore, $\hat{s}^*$ is the size factor estimate for test sample using the required values from training set, $\hat{g}_i$ and $\hat{d}_{ik}$ are estimated parameters taken from training set. A new test sample is assigned to class k which has the highest score in 3.14.

Negative binomial model converges to Poisson model when there is little or no dispersion in data. As $\phi_i \to 0$, we have limiting quantities that $\log(1 + \omega_{i*k}\phi_i) \to 0$ and $\log(1 + \omega_{i*k}\phi_i)^{\phi_i^{-1}} \to \omega_{i*k}$ yielding that

$$\begin{aligned} \log P(y^* = k \mid \mathbf{x}^*) &= \sum_{i=1}^{p} x_i^* \log \hat{d}_{ik} - \sum_{i=1}^{p} \hat{\omega}_{i*k} + \log \hat{\pi}_k + c' \\ &= \sum_{i=1}^{p} x_i^* \log \hat{d}_{ik} - \hat{s}^* \sum_{i=1}^{p} \hat{g}_i \hat{d}_{ik} + \log \hat{\pi}_k + c' \end{aligned} \tag{3.15}$$

which is identical discriminant scores as in 3.9 (31). It is clear from the equation 3.14 that dispersion parameter ϕ_i has an important role in the model. In contrast to Poisson model, a feature having non-zero dispersion ($\phi_i \neq 0$) will be included in the model even if it is not differentially expressed among classes, i.e $d_{ik} = 1$ for all $k = 1, 2, \ldots, K$. Furthermore, the effect of an insignificant feature on the discriminant score will be directly proportional to the magnitude of its dispersion parameter. Hence, another important issue arises that how overdispersion parameter is estimated. In the next subsection, we will cover the estimation method of dispersion parameter in details.

3.2.3. Estimating Dispersion Parameter, ϕ_i

The dispersion parameter ϕ_i indicates how much the variance of a negative binomial distribution exceeds the mean. It can be simply estimated by using maximum likelihood under independence of features and samples. The method-of-moment gives dispersion estimate for probability density function 3.13 as

$$\begin{aligned} \mathrm{Var}\left(X_{ij}\right) &= \mu_{ij} + \mu_{ij}^2 \phi_i, \quad E\left(X_{ij}\right) = \mu_{ij} \\ \phi_i &= \frac{\mathrm{Var}\left(X_{ij}\right) - \mu_{ij}}{\mu_{ij}^2} \end{aligned} \tag{3.16}$$

Note that the dispersion estimates in 3.16 might be a negative value. Maximum likelihood estimation is one of possible solutions for estimating dispersion parameter. However, it is not a robust method when sample size is not large enough in an RNA-seq experiment. Although estimating dispersions as in 3.16 is usually not a problem for large samples, more robust methods are required for better

inferences of differential expression when sample size is small (25, 26). A number of authors studied on alternative ways of estimating dispersion parameter, e.g DESeq (25, 26), edgeR (27, 41), baySeq (46), sSeq (60), etc.

It is reported that maximum likelihood method generally underestimates variance parameters (41). For this reason, direct use of MLEs in 3.16 is not recommended; however, they have been often used as an initial estimate (or initial step) of dispersion parameters in robust methods. Love et al. (26) and Anders and Huber (25) proposed a shrinkage estimation for dispersion parameters in DESeq. They obtained gene-wise MLEs in the first step and fitted a curve to initial estimates using mean-variance relationship. Finally, the estimated dispersions are shrinked towards fitted values. The amount of shrinkage depends on how far initial estimates are from fitted curve. Another approach is edgeR which is proposed by Robinson et al. (27) and based on gene-wise conditional log-likelihoods. The baySeq, on the other hand, iteratively estimates dispersion parameters; hence, requires relatively more time. Among others, we preferred using shrinkage estimator proposed by Yu et al. (60) in sSeq. This method combines the mathematical background of DESeq, edgeR and baySeq; yet, as authors stated that it is simpler, faster and requires less assumptions. The shrunken dispersion estimates are obtained by

$$\tilde{\phi}_i = \delta\xi + (1-\delta)\,\hat{\phi}_i \tag{3.17}$$

where $\hat{\phi}_i$'s are initial dispersion estimates obtained from 3.16. However, negative dispersion estimates are not allowed and the initial estimates are truncated at zero $\hat{\phi}_i = \max(0, \hat{\phi}_i)$. Finally, the initial estimates are shrunken towards a target value ξ using a weight $\delta \in [0,1]$ which is obtained by

$$\delta = \frac{\sum_{i=1}^{p} \left[\hat{\phi}_i - (1/p)\sum_{i=1}^{p}\hat{\phi}_i\right]^2 / (p-1)}{\sum_{i=1}^{p} \left(\hat{\phi}_i - \xi\right)^2 / (p-2)} \tag{3.18}$$

The optimal value of ξ is determined by minimizing the average squared difference between $\hat{\phi}_i$ and $\tilde{\phi}_i$ such that $(1/p)\sum_{i=1}^{p}(\hat{\phi}_i - \tilde{\phi}_i)^2$. As δ approaches 1, initial estimates are forced to be shrunken towards pre-defined target value. Similarly, when $\delta = 0$, no shrinkage is performed on initial estimates $\hat{\phi}_i$. The effect of dispersion parameter is crucial when classifying a new sample. For this reason, results might significantly differ among selected estimation methods. Shrinkage estimator 3.17 performs well comparing to maximum likelihood estimator when

sample size is small. However, when sample size is large enough, the amount of shrinkage on initial estimates are becoming negligible; hence, resulting similar results (26). A comprehensive simulation study is conducted for the comparison of dispersion estimation methods (61). We use shrinkage estimator 3.17 throughout this thesis.

3.2.4. Sparse Negative Binomial Linear Discriminant Analysis

The feature selection procedure for negative binomial model is not clear and simple as in Poisson model. It can be seen from negative binomial discriminant score 3.14 that a feature will be included in the model when $\phi_i \neq 0$ even if $d_{ik} = 0$ for all $k = 1, 2, \ldots, K$. Hence, the model omits a feature if and only if $\phi_i = 0$ and $d_{ik} = 0$ for all $k = 1, 2, \ldots, K$ in the same time. Satisfying these two conditions simultaneously is nearly impossible since $\phi_i \neq 0$ in general.

Dong et al. (31) proposed negative binomial model which uses all features in the model. However, a major number of features are not differentially expressed among classes. Since insignificant features make no contribution to discriminant function, these features should be removed from fitted model. We proposed a two-stage shrinkage algorithm for negative binomial model.

♣ ***Step 1***: In the first step, we follow the shrinkage algorithm 3.10 as in Poisson model and obtained shrunken estimates for differential expression parameter d_{ik}.

♣ ***Step 2***: In the second step, we try to shrink overdispersion parameters towards 0 using shrinkage estimator 3.17. The target value is defined as $\xi = 0$ and the weight δ is automatically calculated within proposed algorithm (60). It is possible to obtain shrinkage estimates for $\tilde{\phi}_i$ which are very close to 0; for example, 0.1, 0.001, etc. For this reason, we proposed truncating such estimates within Step 2 such that

$$\tilde{\phi}_i = \begin{cases} 0, & \text{if} \quad \tilde{\phi}_i \leq \varepsilon \\ \tilde{\phi}_i, & \text{otherwise} \end{cases} \tag{3.19}$$

where ε is a threshold defined by researcher. This is similar to truncating initial estimates of overdispersion parameter at zero, except that shrinkage estimates $\tilde{\phi}_i$ are truncated at ε. Truncating shrinkage estimators at some value is useful when a feature is included in the model due to non-zero overdispersion effect even if it is not differentially expressed among classes.

The amount of sparsity is measured as the percentage of features included in the model. As sparsity approaches zero, the complexity of a model decreases

and less features are included in the model. Because of the fact that the sparsity of a negative binomial model depends on two parameters, ρ and ϕ_i, negative binomial model is generally more complex comparing to Poisson model. The amount of sparsity in a negative binomial model is bounded with

$$S_{max} = \frac{100}{p} \sum_{i=1}^{p} I(\tilde{\phi}_i = 0). \tag{3.20}$$

As $\tilde{\phi}_i \to 0$ for all $i = 1, 2, \ldots, p$, the negative binomial model approximates to Poisson model. Hence, the sparsity measure of each model becomes identical.

3.3. Nearest Shrunken Centroids

In statistical learning, a model with the highest classification accuracy but less complexity is generally desired. However, when $p \gg n$, the complexity of the model drastically increases, and in some cases the convergence problem comes out. As a result, the models which are able to decrease the complexity by selecting a small subset of all features are required. Tibshirani et al. (16) proposed nearest shrunken centroids (NSC) algorithm to overcome high-dimensionality problem. This algorithm is a special case of nearest centroids and diagonal linear discriminant analysis (62, pages: 651 – 654). The aim of NSC algorithm is to keep features in the model which contributes classification the most, likely sparse versions of NBLDA and PLDA. A test sample $\mathbf{x}^* = \{x_1^*, x_2^*, \ldots, x_p^*\}$ is assigned to class k which maximizes the following discriminant function

$$\log P(y^* = k \mid \mathbf{x}^*) = \frac{1}{2}\psi_k(\mathbf{x}^*) - \log\left(\sum_{k=1}^{K} e^{\frac{1}{2}\psi_k(\mathbf{x}^*)}\right) \tag{3.21}$$

where $\psi_k(.)$ is defined as

$$\psi_k\left(\mathbf{x}^*\right) = -\sum_{i=1}^{p} \frac{\left(x_i^* - \bar{x}_{ik}\right)^2}{s_i^2} + 2\log \pi_k \tag{3.22}$$

Here s_i^2 is the pooled variance of i-th gene over K classes, π_k is the prior probability of k-th class and $\bar{x}_{ik} = \sum_{j \in I_k} x_{ij}/n_k$ is the mean (also named as *centroid*) observed counts of i-th feature in class k where n_k is the length of I_k. The vector of indices I_k is defined in the same way of 3.3.

The classification rule in 3.21 uses all features in the model even if a major part has small or no contribution to classification problem. For this reason, a sparse version is proposed for selecting features contributing discriminant function

the most. Let d_{ik} be such that

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{\upsilon_k \left(s_i - s_0\right)} \tag{3.23}$$

where s_0 is a small positive constant which is generally calculated as a median statistic of s_i estimates, $\bar{x}_i$ is the overall centroid of i-th feature and $\upsilon_k = \sqrt{1/n_k - 1/n}$. The shrinkage algorithm aims to shrink gene-wise centorids $\bar{x}_{ik}$ towards 0 by soft thresholding (similar to sparse PLDA algorithm)

$$d_{ik}^{'} = \text{sign}\left(d_{ik}\right)\left(\left|d_{ik}\right| - \rho\right)_+ . \tag{3.24}$$

Here ρ is a threshold parameter to be optimized using , for example, k-fold cross-validation. The shrinkage algorithm in 3.24 focuses on positive side and shrinks any d_{ik} values towards 0 if its absolute value is below threshold. The shrunken centroids $\bar{x}_{ik}^{'}$ are obtained by using 3.24 in 3.23 such that

$$\bar{x}_{ik}^{'} = \bar{x}_i + \upsilon_k \left(s_i + s_0\right) d_{ik}^{'} \tag{3.25}$$

Finally, we replaced the original $\bar{x}_{ik}$ with $\bar{x}_{ik}^{'}$ in 3.22 and obtained discriminant scores from 3.21. Note that a feature satisfying that $d_{ik}^{'} = 0$ for all $k = 1, 2, \ldots, K$ will be removed from model since it does not contribute to discriminant function.

The parameter d_{ik} plays the same role as offset parameter in PLDA algorithm. However, in NSC algorithm, it does not act as an offset parameter but a simple t-statistic. The denominator of d_{ik} is equal to the standard error of the contrast given in the numerator. In conclusion, the class centroids have been shrunken towards overall centroid $\bar{x}_i$ after standardizing observed counts of each class by its standard deviation (16, 48, 62).

3.4. Voom-based Nearest Shrunken Centroids

In this section, we will discuss voom-based nearest shrunken centroids (voomNSC) algorithm as an extension of NSC algorithm. Zararsiz et al. (49) introduced voom (an acronym for 'variance modeling at the observational level') transformation into NSC algorithm in order to improve classification accuracy of NSC algorithm. Voom transformation is proposed by Law et al. (28) around the idea that exploring the mean-variance relationship is more important than specifying the correct underlying distribution of counts. The mean-variance relationship is non-parametrically and also robustly estimated by fitting a locally weighted regression (LOWESS) curve to gene-wise standard deviation of log-cpm

(logarithm of counts per million) values as a function of average log-count (63). The voom transformation is especially suggested for experiments when the sequencing depths are different, i.e unequal library sizes for each sample.

The total number of reads for a given gene depends on not only the expression level of the gene but also the gene length and sequencing depth. We first start by converting read counts of each sample into same scale by using the logarithm of counts per million (log-cpm),

$$z_{ij} = \log_2 \left(\frac{x_{ij} + 0.5}{R_j + 1.0} \times 10^6 \right) \tag{3.26}$$

where $R_j = x_{.j}$ is the library size of j-th sample. The observed counts are shifted by an amount of 0.5 and library sizes by 1.0 to avoid logarithm of zero counts. It also ensures that $(x_{ij} + 0.5)/(R_j + 1.0)$ is strictly within the interval $[0, 1]$.

3.4.1. Estimating Variance of log-cpm Values – Delta Rule

We assume that the observed counts follow negative binomial distribution and the mean-variance relationship is defined as in equation 3.16. When x_{ij} is large, the log-cpm values approximate to

$$z_{ij} \approx \log_2(x_{ij}) - \log_2(R_j) + 6 \log_2(10) \tag{3.27}$$

It also satisfies that $\mathrm{Var}(Z_{ij}) \approx \mathrm{Var}(\log_2(X_{ij}))$. For large values of μ_{ij}, we may define

$$\log_2(x_{ij}) \approx \mu_{ij} + \frac{x_{ij} - \mu_{ij}}{\mu_{ij}} \tag{3.28}$$

where $\mu_{ij} = E(X_{ij})$. Finally, using delta-rule and by Taylor's theorem (64), we obtained

$$\mathrm{Var}(Z_{ij}) \approx \frac{\mathrm{Var}(X_{ij})}{\mu_{ij}^2} = \frac{1}{\mu_{ij}} + \phi_i \tag{3.29}$$

3.4.2. Voom Transformation and Precision Weights

It is possible to explore the effect of several factors and/or covariates on gene expression levels, e.g batch effect, treatment effect, etc. We use following linear model for measuring the effect of covariates on expected log-cpm values

$\mu_{ij}^{(Z)}$ similar to previous studies (9, 65, 66),

$$E(Z_{ij}) = \mu_{ij}^{(Z)} = x_j^T \beta_i \tag{3.30}$$

Here x_j^T is a vector of covariates for j-th sample and β_i is a vector of unknown regression parameters representing the $\log_2$-fold-changes between classes for i-th gene. Note that x is used to represent covariate vectors in 3.30 and should not be considered as mapped read counts.

The linear model 3.30 is fitted to log-cpm values, z_{ij}, for each gene by using ordinary least squares. We obtained fitted values $\hat{\mu}_{ij}^{(Z)}$ and the residual standard deviations $s_{g(e)}$. Finally, the fitted values $\hat{\mu}_{ij}^{(Z)}$ are transformed into fitted mean read counts $\hat{\mu}_{ij}$ as

$$\hat{\mu}_{ij} = \hat{\mu}_{ij}^{(Z)} + \log_2(R_j + 1) - 6\log_2(10) \tag{3.31}$$

We also calculated mean log-cpm values $\bar{z}_i = n^{-1}\sum_{j=1}^{n} z_{ij}$ for each gene and converted to mean log-counts as

$$\tilde{z}_{ij} = \bar{z}_i + \log_2(\tilde{R}_j) - 6\log_2(10) \tag{3.32}$$

where $\tilde{R}_j = \exp[n^{-1}\sum_{j=1}^{n}\log(R_j + 1)]$ is the geometric mean of the shifted library sizes. Next, we obtained a smooth LOWESS curve which shows the mean-variance trend by fitting $s_{g(e)}^{1/2}$ as a function of $\tilde{z}_{ij}$. The mean log-counts $\tilde{z}_{ij}$ are sorted in ascending order and a piecewise linear function $lo(.)$ is obtained by interpolating the LOWESS curve. The function $lo(.)$ is used to obtain predicted variance of log-cpm values z_{ij} as

$$\widehat{\mathrm{Var}}(z_{ij}) = lo\,(\hat{\mu}_{ij})^4 \tag{3.33}$$

Finally, the precision weights are obtained as an inverse value of predicted variance such that $w_{ij} = lo\,(\hat{\mu}_{ij})^{-4}$. We are now able to use log-cpm values and its precision weights as an input to downstream analysis such as differential expression and machine learning.

3.4.3. Extending Nearest Shrunken Centroids to voomNSC

Model training workflow for voomNSC is almost the same with NSC algorithm except that weighted statistics over log-cpm values are used in voomNSC algorithm. Basically, we follow the steps,

1. obtain log-cpm values and replace raw sequencing reads x_{ij} with log-cpm values z_{ij},
2. calculate required centroid statistics from z_{ij} weighting by w_{ij} and use them in replace with corresponding statistics,
3. use log-cpm values and weighted statistics in discriminant function of NSC algorithm to classify new samples.

The log-cpm values and the precision weights are already obtained. We define class-specific weighted gene centroids over log-cpm values by

$$\bar{z}_{ik(w)} = \frac{\sum_{j \in I_k} w_{ij} z_{ij}}{\sum_{j \in I_k} w_{ij}} \tag{3.34}$$

and the weighted overall gene centroids by

$$\bar{z}_{i(w)} = \frac{\sum_{j=1}^{n} w_{ij} z_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{3.35}$$

Furthermore, we define class-specific weighted variances for i-th gene by

$$s^2_{ik(w)} = \frac{\sum_{j \in I_k} w_{ij}}{\left(\sum_{j \in I_k} w_{ij}\right)^2 - \sum_{j \in I_k} w^2_{ij}} \sum_{j \in I_k} w_{ij} \left(z_{ij} - \bar{z}_{ik(w)}\right)^2 \tag{3.36}$$

Finally, the pooled weighted standard deviation of i-th gene is defined by

$$s^2_{i(w)} = \frac{\sum_{k=1}^{K} (n_k - 1) s^2_{ik(w)}}{n - K} \tag{3.37}$$

The weighted statistics which are obtained from equations 3.34 – 3.37 are used within equations 3.23 – 3.25.

3.4.4. Classifying New Samples

A test sample $\mathbf{x}^* = \{x^*_1, x^*_2, \ldots, x^*_p\}$ is assigned to class k using the discriminant function 3.21 of NSC algorithm. However, we replace $\mathbf{x}^*$ with its log-cpm values $\mathbf{z}^*$ using the transformation 3.26. Finally, a log-cpm transformed test sample $\mathbf{z}^*$ is assigned to class k which maximizing the discriminant score $\log P(y^* = k \mid \mathbf{z}^*)$ using the equations 3.21 and 3.22.

3.5. Clustering RNA-Seq Data Using Poisson Dissimilarities

Hierarchical clustering is a popular clustering method for normally distributed data. We compute a distance (or dissimilarity) matrix from data and use it for clustering samples and/or features via hierarchical clustering. Although it is possible to cluster features, we first consider clustering samples using a $n \times n$ dissimilarity matrix. Witten (8) stated that for a normally distributed microarray data, squared Euclidean distance between subjects j and j' are proportional to log-likelihood ratio statistic of testing null hypothesis $H_0 : \mu_{ij} = \mu_{ij'}$, $j = 1, 2, \ldots, p$

$$\sum_{i=1}^{p}\left(x_{ij} - \frac{x_{ij} + x_{ij'}}{2}\right)^2 + \sum_{i=1}^{p}\left(x_{ij'} - \frac{x_{ij} + x_{ij'}}{2}\right)^2 \propto \sum_{i=1}^{p}(x_{ij} - x_{ij'})^2 = \|\mathbf{x}_j - \mathbf{x}_{j'}\|^2 \tag{3.38}$$

assuming that $X_{ij} \sim \text{Normal}(\mu_{ij}, \sigma^2)$ and $X_{ij'} \sim \text{Normal}(\mu_{ij'}, \sigma^2)$. However, mapped read counts from an RNA-Seq experiment does not follow a normal distribution but a Poisson distribution as in 3.1. Therefore, we should modify the statistic 3.38 under Poisson distribution. We first estimate Poisson means μ_{ij} and $\mu_{ij'}$ restricted to $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ using one of size factor estimates in a similar way of classification task. Then, we obtain log-likelihood ratio statistic for null hypothesis $H_0 : d_{ij} = d_{ij'} = 1$, $j = 1, 2, \ldots, p$ as

$$\sum_{i=1}^{p}\left[\hat{\mu}_{ij} + \hat{\mu}_{ij'} - \hat{\mu}_{ij}\hat{d}_{ij} - \hat{\mu}_{ij'}\hat{d}_{ij'} + x_{ij}\log(\hat{d}_{ij}) + x_{ij'}\log(\hat{d}_{ij'})\right] \tag{3.39}$$

Here the Poisson log-likelihood ratio statistic 3.39 becomes infinite when x_{ij} or $x_{ij'}$ is zero. Hence, we obtained modified log-likelihood ratio statistic by using posterior means of d_{ij} and $d_{ij'}$

$$\hat{d}_{ij} = \frac{x_{ij} + \beta}{\hat{\mu}_{ij} + \beta}, \qquad \hat{d}_{ij'} = \frac{x_{ij'} + \beta}{\hat{\mu}_{ij'} + \beta} \tag{3.40}$$

where β is the parameter of Gamma(β, β) priors. We consider 3.40 as a measure of dissimilarity between two observations $\mathbf{x}_j$ and $\mathbf{x}_{j'}$. Finally, a $n \times n$ matrix, called *Poisson dissimilarity matrix*, with elements showing the dissimilarities between pairs (j, j') is obtained. Furthermore, the Poisson dissimilarity matrix is used within hierarchical clustering. This is simply a hierarchical clustering, however, it is called as *Poisson clustering* since distance matrix is obtained from Poisson distribution (8).

3.6. Simulation Study

The classification and clustering accuracies might be affected by several factors such as sample size, number of features, overdispersion, number of classes, etc. We conducted a comprehensive simulation study in order to explore the effect of each factor under different scenarios (Figure 3.2). The simulation study contains all combinations of

- number of samples: $n = \{40, 60, 80, 100\}$,
- number of features: $p = \{500, 1000, 2000\}$,
- number of classes: $k = \{2, 3\}$,
- overdispersion: $\phi = \{0.01, 0.1, 1\}$ which corresponds to *low*, *moderate* and *high* overdispersions respectively, and
- differential expression rate: $\varrho = \{0.01, 0.05, 0.10\}$.

There are 216 combinations to be evaluated in our simulation setup. We repeated whole process (Figure 3.2) 100 times for each combination yielding that $21,600$ simulated data sets were analysed by using selected classification and clustering algorithms. In **data generation** step, we randomly generated count data from negative binomial distribution which is parameterized as in equation 3.12. We generated offset parameters from lognormal distribution such that $\log(d_{ik})$ is normally distributed with mean 0 and standard deviation 0.5. Gene totals, g_i, are generated from exponential distribution with a mean of 25, and size factors are generated from uniform distribution within the interval $[0.2, 2.2]$. Finally, we used generated size factors, gene totals and offset parameters to obtain mean of negative binomial distribution. We also generated a test set by using the same distributional parameters except that size factors are different for test set samples. Another alternative is to split generated data set into two parts as training and test sets according to a researcher defined splitting ratio, e.g 70% for training and 30% for test set. The amount of training and test set ratio is an important criteria since it greatly affects model accuracies, especially when sample size is small. In our simulation setup, we used an independent test set rather than splitting data into two parts.

In **pre-filtering step**, we prepared data for downstream analysis by removing genes/samples with low quality. First, we performed *near-zero variance filtering* to remove features with zero or very low variance. We also filtered least abundant features whose total mapped read counts is less than 50 as similar strategy to Leidinger et al. (67). In practice, thousands of genes or gene regions (e.g

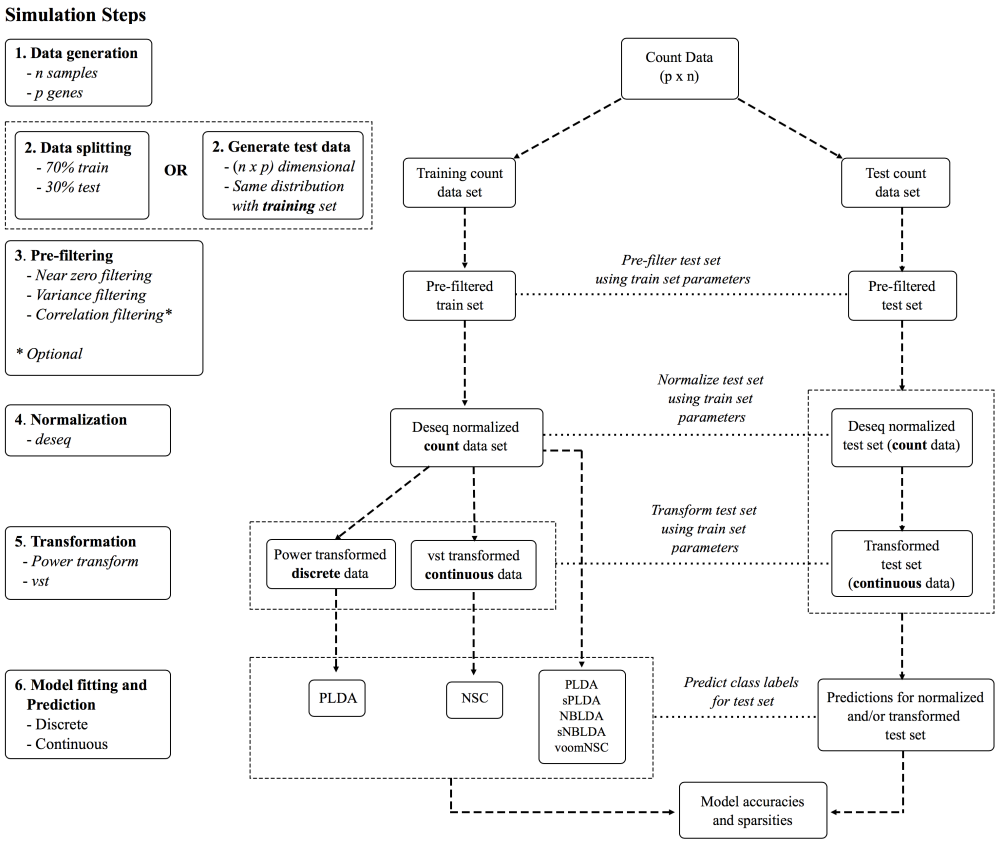


Figure 3.2. A work flow for simulation study – classification.

2,000 miRNA or 20,531 mRNA, etc.) are sequenced in an RNA-seq experiment. However, only a small part of all features, say 20 or 50 for example, is expected to be differentially expressed among classes. In order to reflect a real RNA-Seq experiment, we generated $10,000$ features and only a small amount of these features is set to be differentially expressed. The amount of differential expression rate is controlled by ϱ as 1%, 5% and 10%. Among $10,000$ features, we selected 500, $1,000$ and $2,000$ features by maximum variance filtering[1] and used them for both clustering and classification tasks. A subset of all features are determined by sorting variances in decreasing order due to computational cost of analysing $10,000$ features at once.

Pre-filtered data is then normalized by using *deseq median ratio* in **normalization step**. Other normalization methods were not performed since the comparison of different normalization methods is out of scope. After normalization, data set is transformed by using *power* and *vst* transformations in **trans-**

[1] Another alternative is to select a subset of features by differential expression analysis. We discussed this strategy in 'Discussion' section.

formation step.

Finally, in **model fitting and prediction step**, we fit power transformed data to PLDA, vst transformed data to continuous classifiers and deseq normalized data to voomNSC and sparse/non-sparse versions of PLDA and NBLDA. Even we took normalized data as an input to voomNSC, normalized counts are transformed using voom transformation within voomNSC algorithm. We omitted *rlog* transformation due to several reasons: (i) it is computationally intensive, (ii) rlog and vst transforms perform very similar as sample size increases, and (ii) comparison between different transformation techniques is not among aims of this study.

3.7. Evaluation Process of Model Accuracies

In a classification problem, the predicted and actual class labels can be summarized on a cross table. For example, Table 3.1 shows a 2-by-2 classification table of a binary classification problem where predicted class labels are represented in the rows. A popular metric calculated from classification table for assessing model performance is overall accuracy. However, this measure should be preferred when sample sizes for each class are balanced. Otherwise, it may lead to biased conclusions. When classes are imbalanced, balanced accuracy should be preferred in place of overall accuracy. Beside overall accuracy, other performance measures such as sensitivity, specificity, positive and negative predictive values can be considered for a better conclusion on model performance.

Table 3.1. A classification table (confusion matrix) for a binary classification problem.

	Actual		
Predicted	A	B	Total
A	n_{11}	n_{12}	$n_{1.}$
B	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n = n_{..}$

The selected performance measures for a binary classification problem are given

in equations 3.41 assuming that class A is reference category

$$Sensitivity = Se\,(\mathrm{A}) = n_{11} \,/\, n_{.1} \tag{3.41a}$$

$$Specificity = Sp\,(\mathrm{B}) = n_{22} \,/\, n_{.2} \tag{3.41b}$$

$$Accuracy = ACC = (n_{11} + n_{22}) \,/\, n \tag{3.41c}$$

$$Balanced\ Accucary = bACC = \frac{Sensitivity + Specificity}{2} \tag{3.41d}$$

where $Se\,()$ and $Sp\,()$ are functions which return sensitivity and specificity measures of a given class. Classes for each simulated data set were balanced in our simulation setup. Hence, both accuracy and balanced accuracy measures lead to similar conclusion in this study. We compared model performances by using accuracy measure as in equation 3.41c.

For multi-class classification problems, it is not possible to obtain single value for sensitivity and specificity measures. However, one may obtain sensitivity and specificity measures by using *one versus all* strategy. Table 3.2 gives a confusion matrix for 3-by-3 classification problem. Here, one of categories is considered as reference, and its sensitivity and specificity measures versus all other categories are calculated. Overall accuracy, on the other hand, is easily calculated by using diagonal elements of confusion matrix of multi-class problem.

Table 3.2. A classification table (confusion matrix) for a multi-class classification problem – 3-by-3 table.

	Actual			
Predicted	A	B	C	Total
A	n_{11}	n_{12}	n_{13}	$n_{1.}$
B	n_{21}	n_{22}	n_{23}	$n_{2.}$
C	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n = n_{..}$

The performance measures for Table 3.2 when class A is considered as

reference category are

$$Se\,(\mathrm{A}) = n_{11} \,/\, n_{.1} \tag{3.42a}$$

$$Sp\,(\mathrm{A}') = (n_{22} + n_{23} + n_{32} + n_{33}) \,/\, (n_{.2} + n_{.3}) \tag{3.42b}$$

$$ACC = (n_{11} + n_{22} + n_{33}) \,/\, n \tag{3.42c}$$

$$bACC = \frac{Se\,(\mathrm{A}) + Sp\,(\mathrm{A}')}{2} \tag{3.42d}$$

where $Sp\,()$ is calculated by collapsing all categories other than reference category. We may calculate performance measures for remaining classes in the same way that we performed for class A as in equations 3.42. In addition to overall accuracy, we calculated sparsity[2] measure of sparse classifiers, and compared sparse models with each other and also with its non-sparse version.

3.8. Availability of Proposed Algorithms

The proposed algorithms and other classification and clustering tasks are available through R[3] programming language. R is a free and open source platform developed for statistical analysis and graphics. The statistical methods and graphics are included in R via different *packages* (also called as *libraries*). Hence, researchers can extend R using different packages from The Comprehensive R Archive Network[4] (CRAN) and Bioconductor Network[5].

We developed R packages **MLSeq** (30) through Bioconductor and **NBLDA** (68) through CRAN for implementing proposed algorithms and simulation study in this thesis. Although MLSeq is specifically developed for classification of RNA-Seq data, NBLDA can be used for classifying count data from any field. The NBLDA package is included in MLSeq package, hence, there is no need to install NBLDA package if MLSeq is already installed in R. MLSeq can be installed to R by running following codes in R console.

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("MLSeq")
```

[2] Sparsity is measured as the percentage of features included in the model.

[3] http://www.cran.r-project.org

[4] https://cran.r-project.org/mirrors.html

[5] https://bioconductor.org

3.9. Real Data Sets

In addition to simulation study, we performed clustering and classification tasks on real RNA-Seq data sets (29, 49) . The first data set is *cervical cancer data* which was collected by Witten et al. (57). The authors sequenced miRNAs from 58 human cervical tissue samples (29 tumor and 29 matched control) using Solexa/Illumina platform. The aim of this study was to detect differentially expressed genes between healthy and tumor samples, and to detect novel miRNAs associated with cervical cancer. We used a count matrix in our analysis which consists of 58 samples and 714 miRNAs.

The second data set is *Alzheimer data* which was collected by Leidinger et al. (67). This data set consists of sequencing reads of $2,801$ miRNAs obtained from blood samples of 48 alzheimer patients and 22 age-matched controls. Alzheimer patients were selected among subjects which were diagnosed through several tests including Alzheimer Disease Assessment Scale-cognitive subscale (ADAS-Cog), Wechsler Memory Scale (WMS), Mini-Mental State Exam (MMSE) and Clinical Dementia Rating (CDR). The authors aimed to explore potential miRNAs which are responsible for alzheimer disease and related neurological disorders.

Renal cell carcinoma (RCC) data is the third and last data set that we used in our analysis. This data set is downloaded from The Cancer Genome Atlas (TCGA) which is a comprehensive platform for researchers. It is possible to explore and download data sets among thousands of publicly available sources by using TCGA platform (69). We dowloaded sequencing reads of $20,531$ known human RNAs belonging to $1,020$ RCC patients. The patients were splitted into the three most common sub-categories (70); kidney renal papillary cell (KIRP), kidney renal clear cell (KIRC) and kidney chromophobe carcinomas (KICH) with sample sizes 606, 323 and 91, respectively.

We also used *lung cancer* data which is downloaded from TCGA platform. This data set contains mapped read counts of $20,531$ known human RNAs belonging to $1,128$ lung cancer patients. The patients were diagnosed into two distinct classes of lung cancer which are lung adenocarcinoma (LUAD) and lung squamous cell with carcinoma (LUSC) with sample sizes 576 and 552, respectively. We did not use lung cancer data for classification purpose; however, it is used to generate mean-variance trend plots as given in section 2.4 (Figure 2.2).

4. RESULTS

In this chapter, we reported the results for both simulation study and real data sets. All the analyses were performed in R programming language version 3.5.0. We performed classification analysis for both simulated and real data, however, clustering analysis were performed only for real data sets. We used three RNA-Seq data sets as described in section 3.9 except *lung cancer* data. Results for simulation study and real data are given in details in sections 4.1 and 4.2 respectively. Total number of features for real data sets after pre-filtering steps were below 1000 for two data sets except kidney cancer data. For this reason, we were not able to compare the effect of number of selected features on classification accuracy. Hence, we selected all features for data sets whose number of features were below 1000, and top 2000 features were selected by maximum variance filtering for kidney cancer data.

In clustering part, we reported the results for real data sets because we did not performed a simulation study for clustering. We followed the same pre-filtering steps for clustering and filtered features were used to build clustering models. Samples are clustered using k-means, hierarchical, negative binomial and Poisson clustering algorithms. Transformed values were used within hierarchical and k-means clustering while normalized counts were used within Poisson and negative binomial clustering. Raw counts were transformed using vst, log-cpm and logarithm of deseq normalized counts. Clustering performances were evaluated using adjusted Rand index and average silhouette measure.

4.1. Simulation Results

Simulation results are given in Figures 4.1 – 4.6 for number of classes 2 and 3, and differential expression rates 1%, 5% and 10%. The number of samples are presented in each row while overdispersion parameters are presented in columns on each figure. Hence, each panel of figures corresponds to a sample size overdispersion combination defined as 'sample size:overdispersion'. For example, top left panel of Figure 4.1 shows the results when sample size is 40 and overdispersion is 1 which is shown as 40:1. The amount of overdispersion descreases as moving from left to right, and sample size increases as moving from top to bottom. Each method is given on x-axis where the number '2' in x-axis labels represents power transformation for PLDA and NBLDA algorithms, and variance stabilizing transformation for NSC algorithm. Test set accuracies are shown with error bars and the effect of selected number of features are presented with different line types

within each classifier. We truncated upper limits at 1 and lower limits at 0 if upper and/or lower limits exceeds the interval $[0, 1]$.

The amount of differential expression rate, sample size and number of selected features are positively correlated with model accuracies. As differential expression rate increases, the number of features which are differentially expressed between classes also increases. As a result, classifiers are able to select differentially expressed features and better discriminates between classes. Since there are thousands of features in simulated data sets, we selected a subset of all features due to computational cost. As the number of selected features for each subset increases, model accuracies also increases because it is more likely to include more differentially expressed features in a subset when subset sizes increases. Sample size for training set also has positive effect on model performances. As sample size of training and test set increases, model performances also increases. However, its effect on model performances are negligible for moderately and slightly overdispersed data sets. Furthermore, the effect of sample size on model performances decreases as differential expression rate increases. Results showed that increasing sample size when differential expression rate is low has greater impact on model performances compared to high differential expression rate.

The number of classes and amount of overdispersion, on the other hand, are negatively correlated with model accuracies. Test set accuracies decreases as the number of classes increases since classifying (or clustering) two classes is easier and simpler. Overdispersion has a great effect on model accuracies. Although classification models give very similar results when data set is moderately or slightly overdispersed, test set accuracies significantly decreases when data is highly overdispersed (Figures 4.1 and 4.4).

We compared 10 different classifiers on each data set. These classifiers include discrete classifiers (e.g PLDA and NBLDA) and continuous classifiers fitted to transformed counts (e.g NSC and voomNSC). In general, discrete classifiers performed better comparing to continuous classifiers. Among discrete classifiers, NBLDA performed better than PLDA when data is highly overdispersed. This result is as expected since NBLDA takes overdispersion effect into account. However, as sample size and differential expression rate increases, NBLDA and PLDA gives very similar results even for highly overdispersed data. For example, NBLDA and PLDA performs similar for simulation combination '100:1' when differential expression rate is 10% (Figure 4.3). However, NBLDA performs better for the same simulation combination when differential expression rate is 1% (Figure 4.1). This might be due to number of differentially expressed features that

are included in the model. When differential expression rate is 1%, 100 out of 10000 features are expected to be differentially expressed between classes. However, each model includes at least 500 features in the model where a major part of all features are redundant for classification, i.e features that do not contribute to discrimination. For this reason, it is important to remove redundant features and decrease model complexity in order to obtain better model performance. Sparse classifiers are able to select an optimal feature subset by using built-in variable selection criteria. It can be seen from Figures 4.1 and 4.4 that sparse versions of NBLDA and PLDA performed better. Finally, a power transformation is applied on raw counts and model performances were slightly increased (see NBDLA2 and PLDA2 in figures). Although power transformation and sparse classifiers increase model performances for both NBLDA and PLDA, using sparse classifiers rather than power transformation has better effect on model performances. Performance of microarray based classifiers were slightly below discrete classifiers. Among these classifiers NSC and vNSC performed similar.

In conclusion, classification models was comparable in terms of prediction accuracy when differential expression rate was low and overdispersion was high. However, all models performed similar as data became less overdispersed and more features were differentially expressed (Figures 4.2 and 4.5).

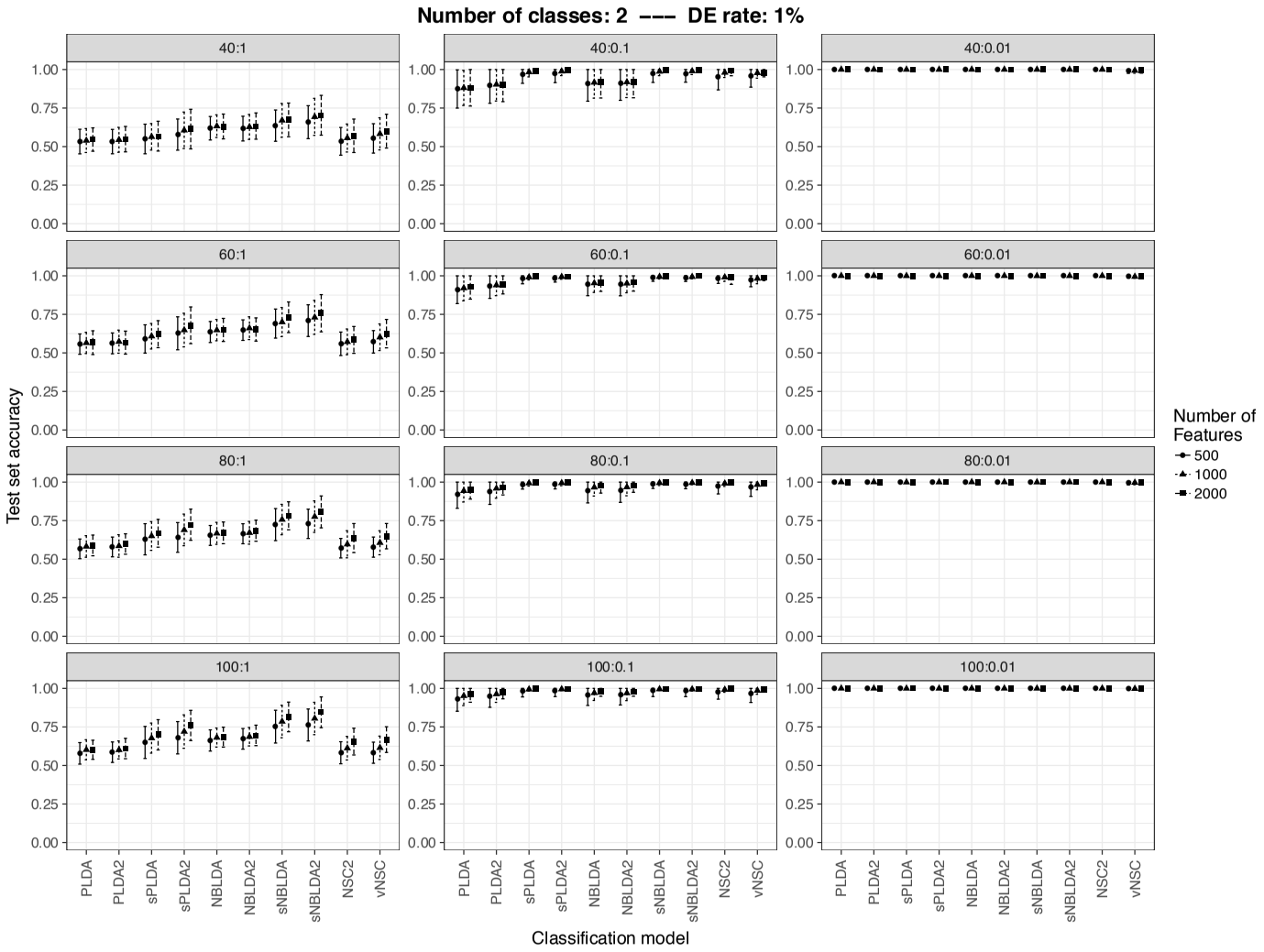


Figure 4.1. Simulation results – Number of groups: 2, Differential expression rate: 1%

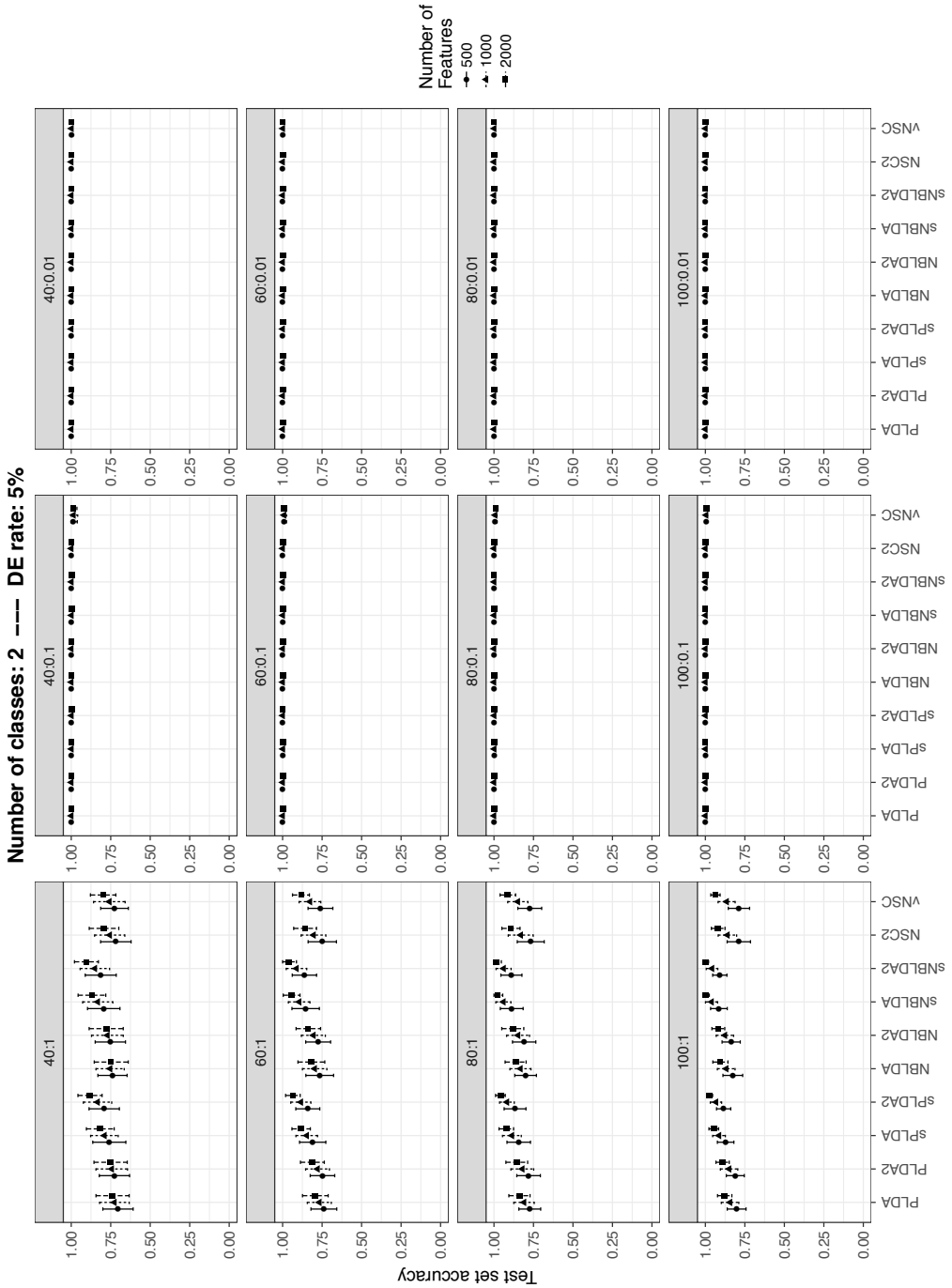


Figure 4.2. Simulation results – Number of groups: 2, Differential expression rate: 5%

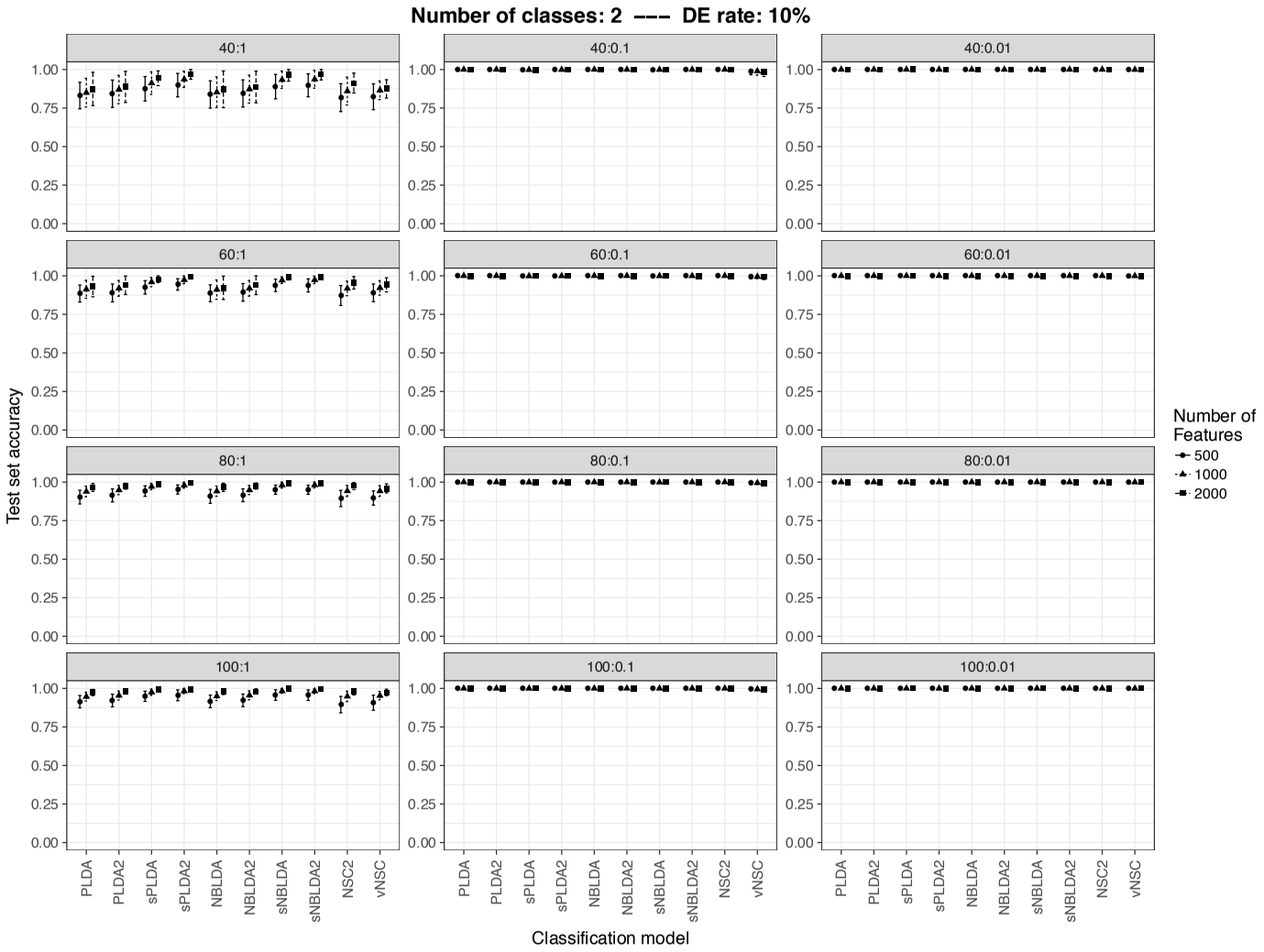


Figure 4.3. Simulation results – Number of groups: 2, Differential expression rate: 10%

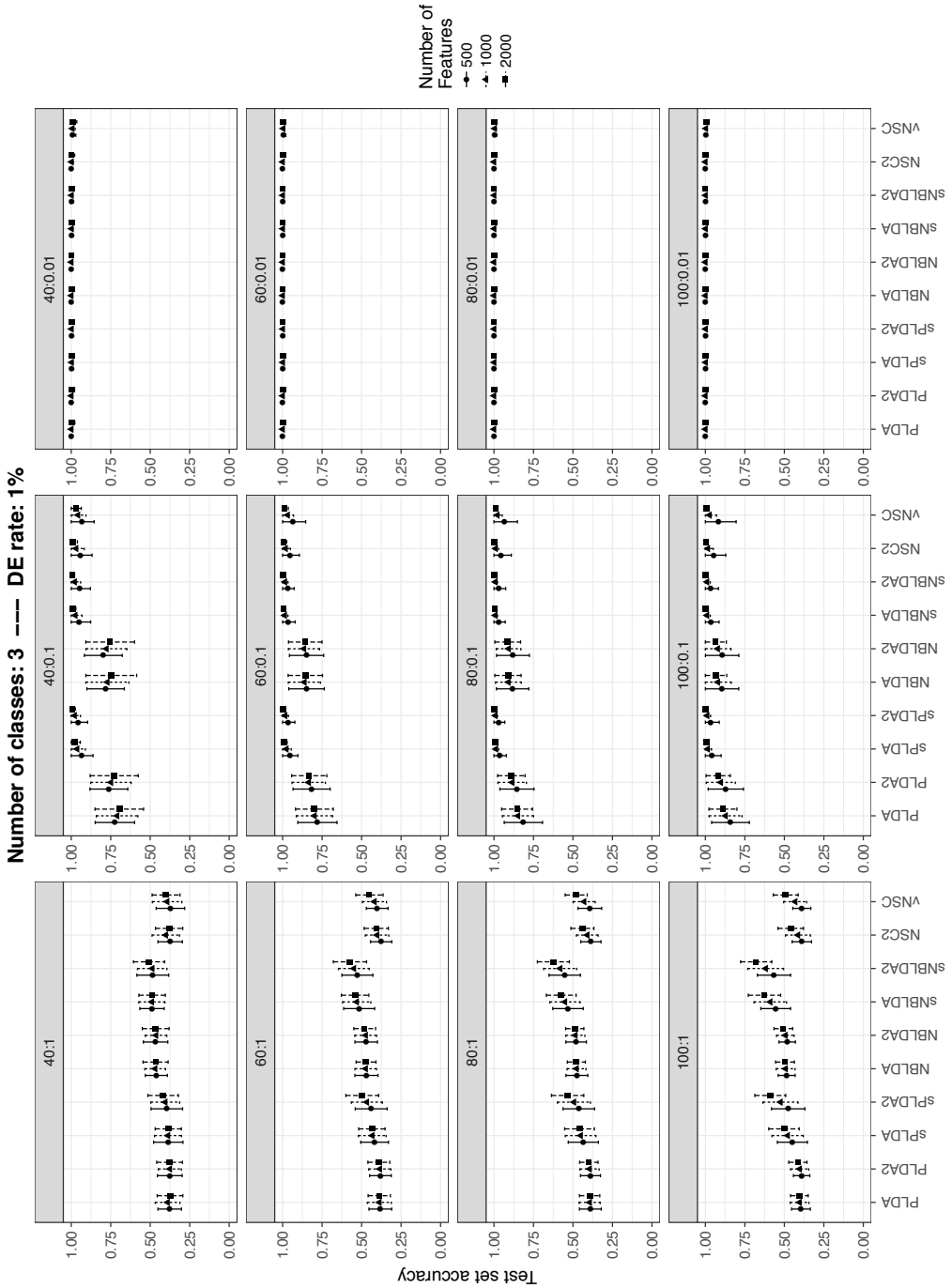


Figure 4.4. Simulation results – Number of groups: 3, Differential expression rate: 1%

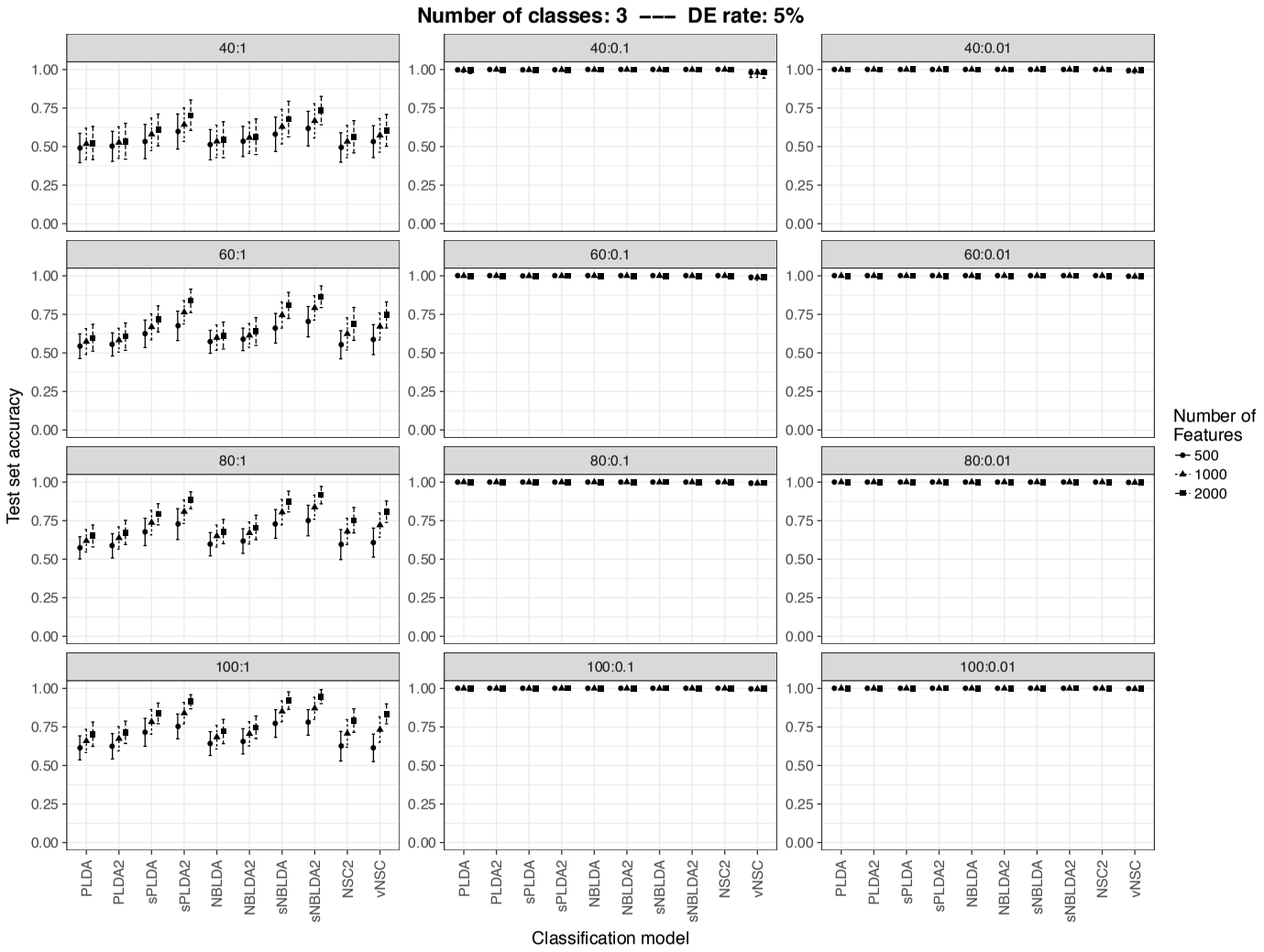


Figure 4.5. Simulation results – Number of groups: 3, Differential expression rate: 5%

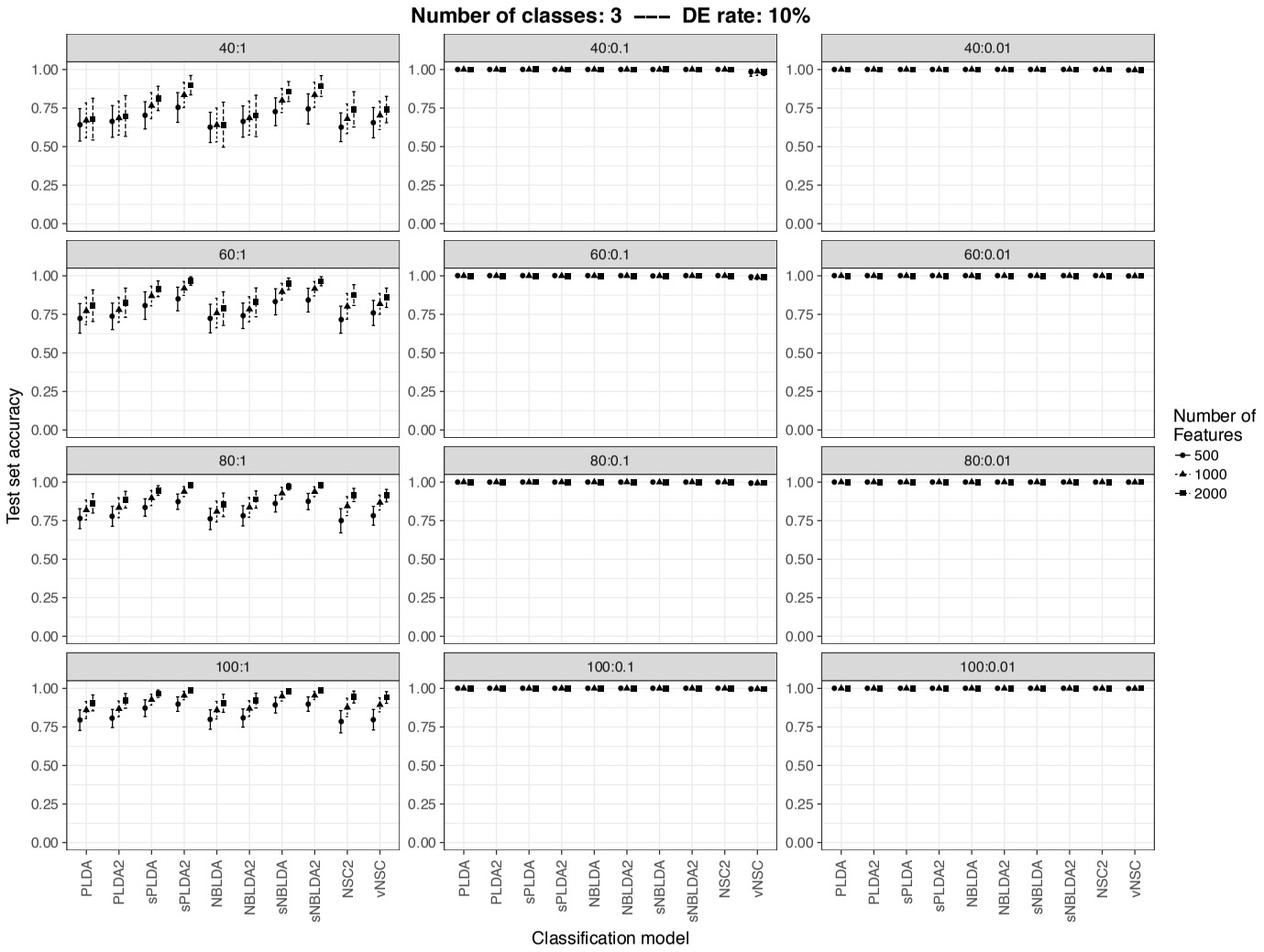


Figure 4.6. Simulation results – Number of groups: 3, Differential expression rate: 10%

4.2. Real Data Results

Real data sets were analysed using the same workflow of simulation study as given in section 3.6 (Figure 3.2). First, we split data into two parts as 70% and 30% for *train* and *test* sets. Next, raw counts were pre-filtered using minimum count and near zero variance filtering, normalized and transformed using the similar methods as in simulation study. Threshold is selected as 50 for minimum count filtering, and features whose total mapped read counts are below threshold were removed. Finally, we fit pre-filtered, normalized and transformed data to selected classifiers and obtained test set accuracies. We repeated whole process 100 times and evaluated mean accuracies and sparsities for each model.

Overdispersions are estimated from complete data (i.e before splitting data as test and train sets) using the mean-variance relationship of negative binomial distribution. The total number of features for cervical cancer, alzheimer disease and kidney cancer data sets were 714, 2801 and 20531, and the total number of features after pre-filtering were 487, 703 and 19305 respectively. An important part of features for cervical and alzheimer data is removed in pre-filtering step. We used normalized counts in order to remove effect of sequencing depth, and the gene-wise overdispersions are estimated from normalized counts. We also calculated the percentage of features in each data set whose overdispersion estimates exceed 1, i.e the threshold level for high overdispersion. Figure 4.7 shows the distribution of estimated gene-wise overdispersions for each data sets after pre-filtering and normalization. This figure is generated after upper 2.5% of the estimated overdispersions are trimmed since few extreme overdispersion estimates yields ugly scaled plots on x-axis. Results showed that 89.1%, 22.1% and 41.8% of all features have highly overdispersed read counts for cervical, alzheimer and kidney data sets respectively. We can say that cervical cancer data is the most overdispersed while alzheimer data is the least overdispersed.

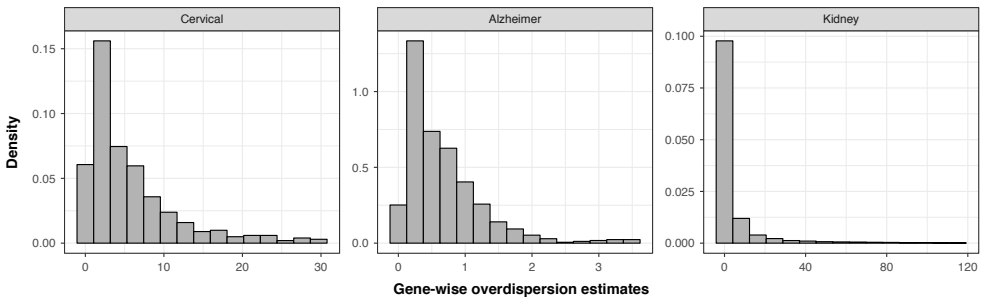


Figure 4.7. Gene-wise overdispersion estimates for real data sets

Classification results for cervical cancer data is given in Figure 4.8. Accord-

ing to the results, negative binomial classifiers performed the best. Classification accuracies increased when power transformation was performed on normalized counts for PLDA and NBLDA classifiers. However, the effect of power transformation was greater for PLDA than NBLDA.

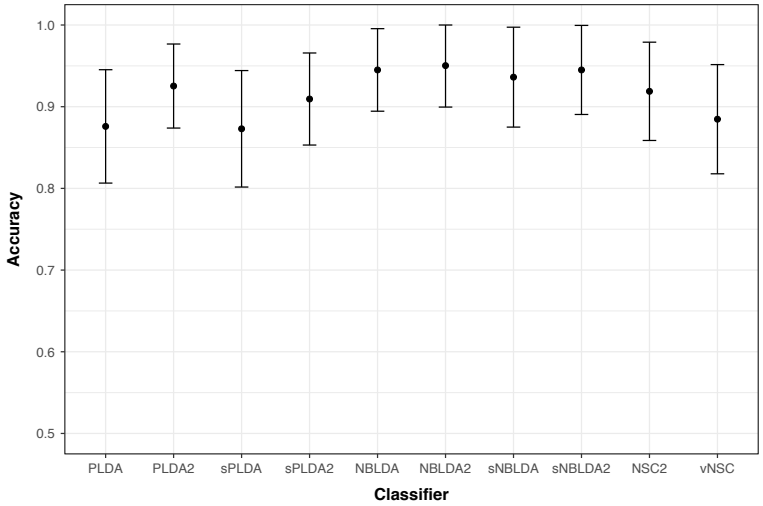


Figure 4.8. Classification results for cervial cancer data

In addition to test set accuracies, we compared the sparsity of sparse classifiers. The distribution of selected features for each sparse classifier is given in Figure 4.9 over 100 repeats. Power transformed classifiers were more sparse for PLDA and NBLDA. Among sparse classifiers, voom-based NSC (vNSC) was the most sparse classifier. When classifiers were evaluated according to accuracy and sparsity measures, we found that power transformed sparse NBLDA was the best classifier for cervical cancer data.

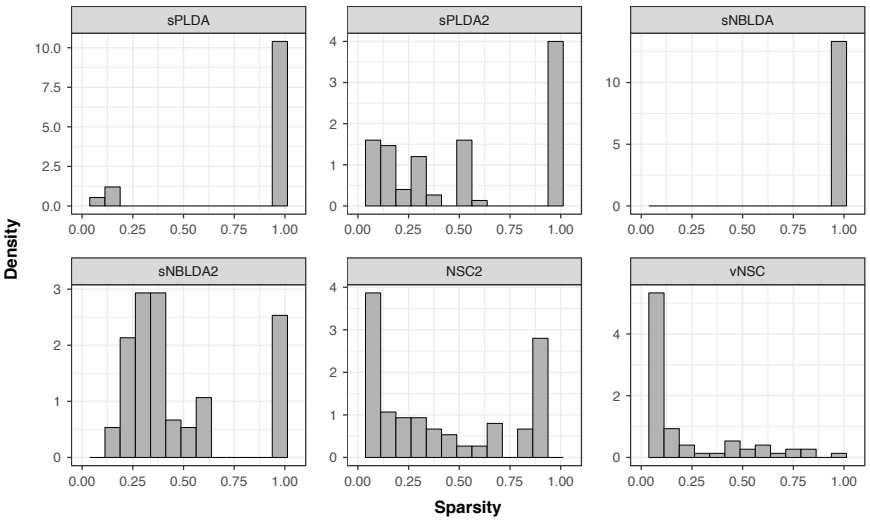


Figure 4.9. Sparsity results for cervical cancer data

Classification results for alzheimer disease data are given in Figures 4.10 and 4.11. Although alzheimer data is the least overdispersed data set among others, PLDA gave the worst classification accuracy unless power transformation was performed. When power transformation was performed, PLDA and NBLDA classifiers performed very similar. Furthermore, NSC and vNSC classifiers were also performed similar to NBLDA classifiers. Figure 4.11 showed that voom and variance stabilizing transformation based classifiers were more sparse than count based classifiers. Among sparse classifiers, vNSC was the most sparse classifier for alzheimer disease data.

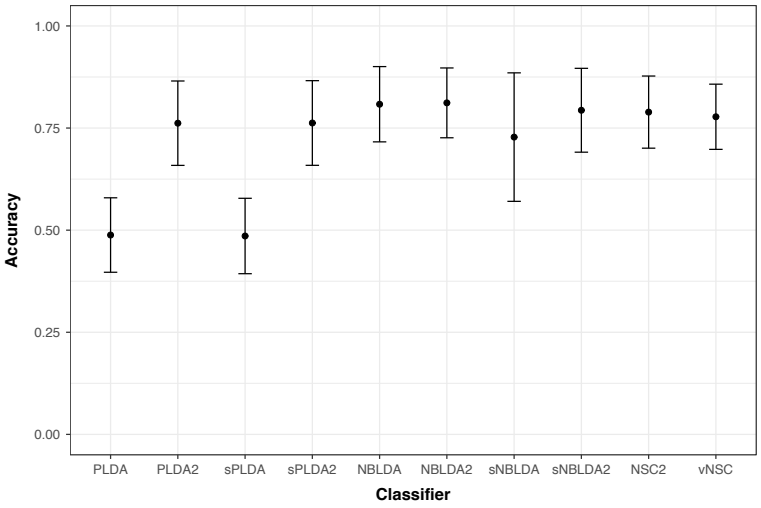


Figure 4.10. Classification results for Alzheimer disease data

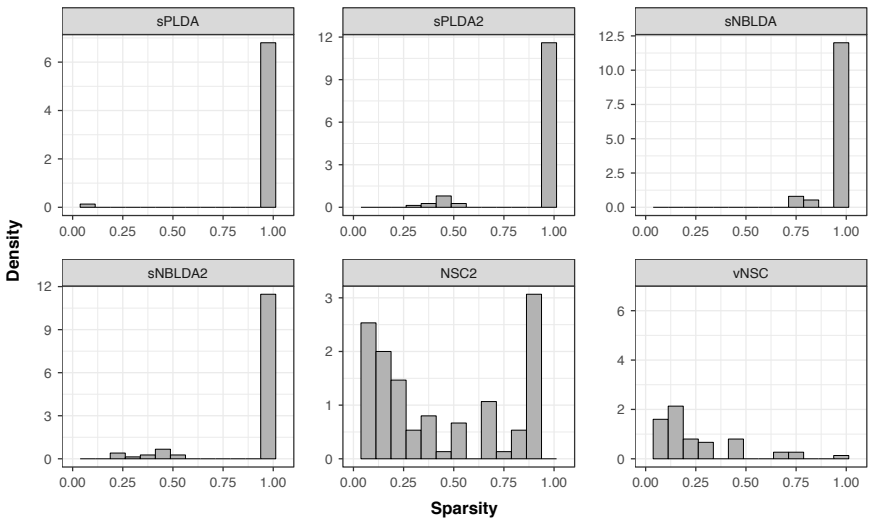


Figure 4.11. Sparsity results for Alzheimer disease data

In kidney cancer (renal cell carcinoma) data set, negative binomial models performed the best while vNSC performed the worst. It can be seen from Figure

4.12 that NBLDA classifiers were better than PLDA classifiers as in alzheimer disease and cervical cancer data sets. Sparse classifiers, except vNSC, were not able to select a subset of all features for kidney cancer data (Figure 4.13). vNSC algorithm was used approximately 6% of all features in the classification; however, test set accuracy was lower comparing to other classifiers.

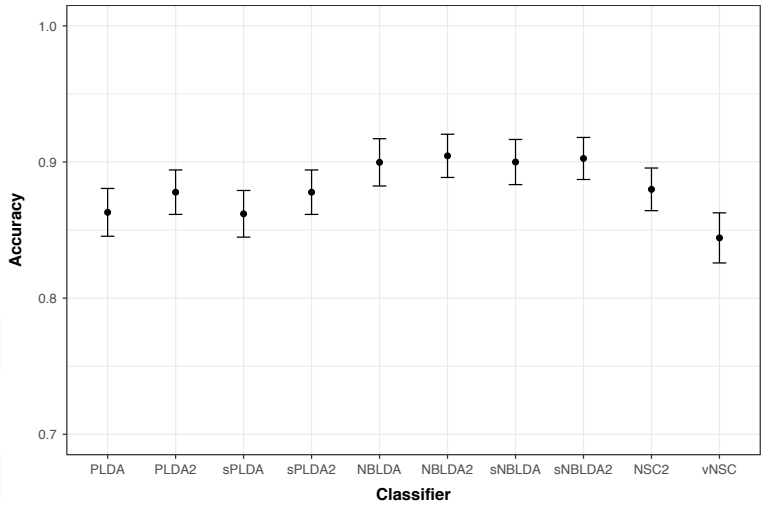


Figure 4.12. Classification results for kidney cancer data

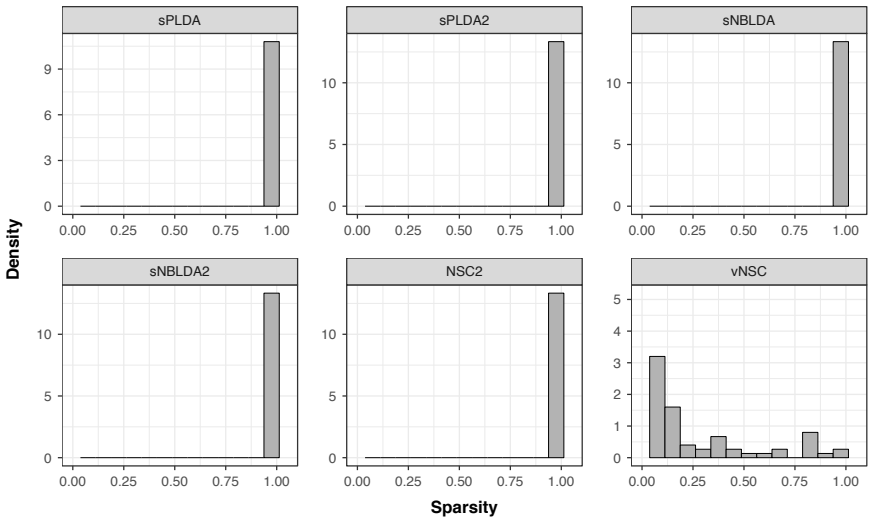


Figure 4.13. Sparsity results for kidney cancer data

We summarized classification results of all classifiers for three real data sets in Table 4.1. For cervical data set, sparse NBLDA classifier with power transformation was the best classifier with a classification accuracy of 0.94, and included only 36.3% of all features after pre-filtering. For alzheimer disease data, NSC and NBLDA algorithms performed very similar. Finally, for kidney cancer data set, negative binomial models performed the best. We also reported feature

selection results and class information for each data set in this table. On the average, 455.2, 634 and 2000 features remained after pre-filtering step for cervical, alzheimer and kidney data sets respectively. Furthermore, alzheimer and kidney data sets were imbalanced within classes. Class imbalance ratios were 2.18 for alzheimer data and 6.62 for kidney data.

Table 4.1. Classification results for real data sets

	Cervical		Alzheimer		Kidney	
Number of features						
Raw data	714		2801		20531	
Pre-filtered (Avg.)	455.2		634		2000	
Class sizes	29/29		22/48		91/323/602	
Class ratios	1:1		1:2.18		1:3.55:6.62	
Models	Accuracy	Sparsity	Accuracy	Sparsity	Accuracy	Sparsity
PLDA	0.8759		0.4880		0.8630	
PLDA2	0.9253		0.7620		0.8778	
sPLDA	0.8729	1.0000	0.4860	1.0000	0.8619	1.0000
sPLDA2	0.9094	0.2990	0.7620	1.0000	0.8778	1.0000
NBLDA	0.9400		0.7830		0.8997	
NBLDA2	0.9453		**0.7870**		**0.9045**	
sNBLDA	0.9312	1.0000	0.7030	1.0000	0.8999	1.0000
sNBLDA2	**0.9400**	**0.3630**	0.7690	1.0000	0.9025	1.0000
NSC2	0.9189	0.2850	**0.7890**	**0.3120**	0.8800	0.9840
vNSC	0.8847	0.0543	0.7780	0.0375	0.8442	0.0600

* Accuracies are given as mean and sparsities are given as median of 100 repeats.

Clustering results of each real data set are given in Table 4.2. Negative binomial and Poisson clustering performed better for cervical and kidney data because these data sets are highly overdispersed (Figure 4.7). However, alzheimer data is poorly clustered using clustering algorithms based on discrete distributions. Hierarchical clustering on vst and log2-normalized counts was the best model for alzheimer data. Clustering performances may be affected by not only overdispersion but also selected linkage method, number of samples, number of features, distance measure etc. Hierarchical, negative binomial and Poisson clustering results in Table 4.2 were obtained using average linkage method, and euclidean distance is used as dissimilarity measure.

According to adjusted Rand index and silhouette measure, clusters for alzheimer data was not well separated. However, clusters for cervical and kidney data sets were quite separable. This can be graphically shown by using

Table 4.2. Clustering results for real data sets

	Cervical		Alzheimer		Kidney	
Number of features						
Overall	714		2801		20531	
Pre-filtered	488		703		2000	
Class sizes	29/29		22/48		602/323/91	
Models	Rand	Silhouette	Rand	Silhouette.	Rand	Silhouette
K-means						
log2-normalized	**0.6793**	0.1550	0.1490	0.1569	0.2726	0.1153
vst	**0.6793**	0.1537	0.1248	0.1670	**0.6250**	0.2463
log-cpm	0.0596	0.3257	0.0031	0.2232	0.5783	0.2118
Hierarchical						
log2-normalized	0.5680	0.1386	**0.1691**	0.1622	0.5888	0.2353
vst	0.5680	0.1367	**0.1691**	0.1682	**0.6200**	0.2448
log-cpm	0.0094	0.2701	0.0145	0.2344	0.5970	0.2175
Neg. Bin. (edgeR)	0.6224	0.1321	0.1493	0.1513	0.5910	0.2405
Poisson						
Transformed	0.2575	0.2517	0.0348	0.4065	0.5853	**0.3937**
Untransfromed	0.5683	**0.3704**	0.036	**0.6922**	**0.6173**	0.3368

* Adjusted Rand Index and Average Silhouette measures are reported.

2-dimensional principal component plots for vst transformed values. Figure 4.14 also showed that two clusters, diseased (AD) and control groups (C), are not well separated for alzheimer data set. However, two distinct clusters can be seen for cervical cancer data set. There were three subgroups for kidney cancer data set where KIRP was the largest and KICH the smallest subgroups in terms of sample sizes. Although there were three distinct clusters for kidney data (Figure 4.14c), an important part of samples from KIRP group was incorrectly clustered with KICH group. In conclusion, KIRP and KICH clusters were moderately separable while KIRC cluster was quite separated from other clusters.

Another useful graphic which can be used to evaluate clustering performances is silhouette graph. This graph shows silhouette measures of each sample which represents how well each object assigned to its cluster. Figure 4.15 shows silhouette graphs for vst transformed values of cervical and alzheimer data sets. We use silhouette graph to visually assess cluster quality. A cluster is well defined if silhouette measures of subjects belonging to this cluster are large. Although we obtained high adjusted rand index for cervical than alzheimer data, within cluster qualities was slightly higher for alzheimer data. Hence, adjusted rand index and silhouette measures should be evaluated together. Adjusted Rand index can be

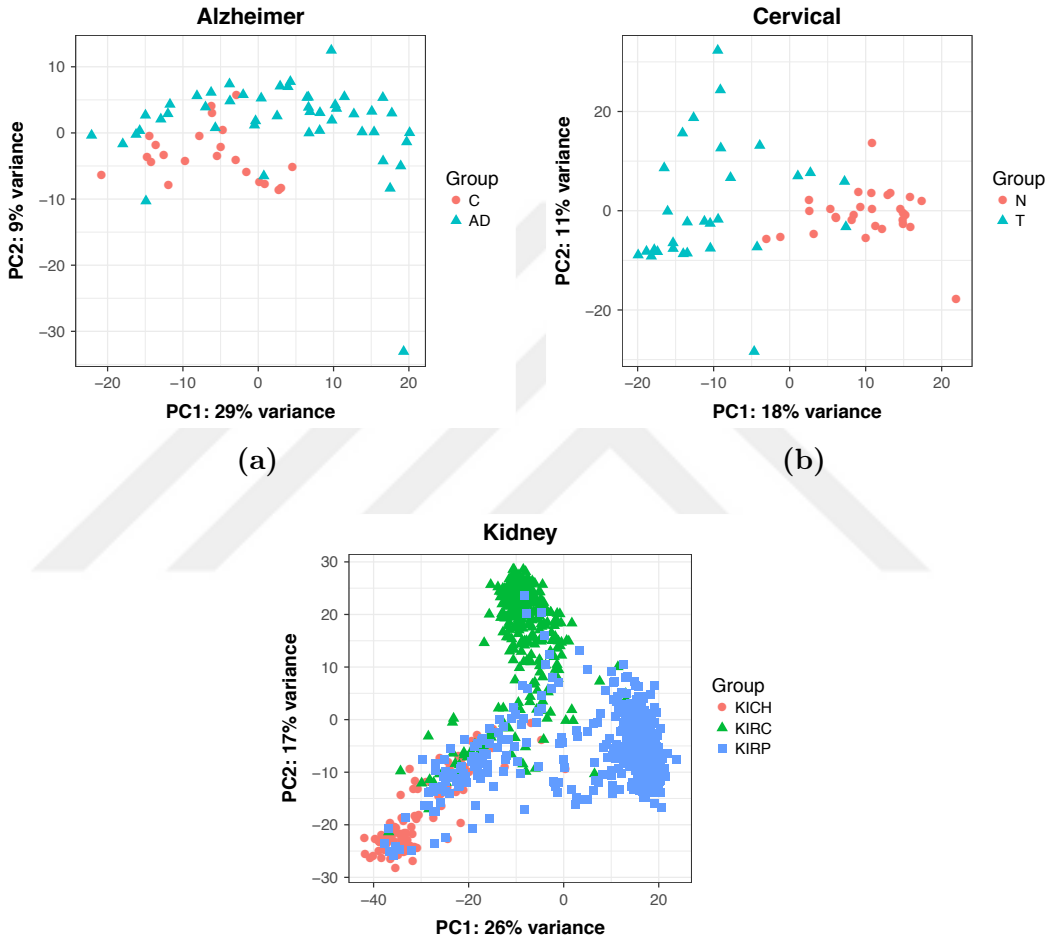
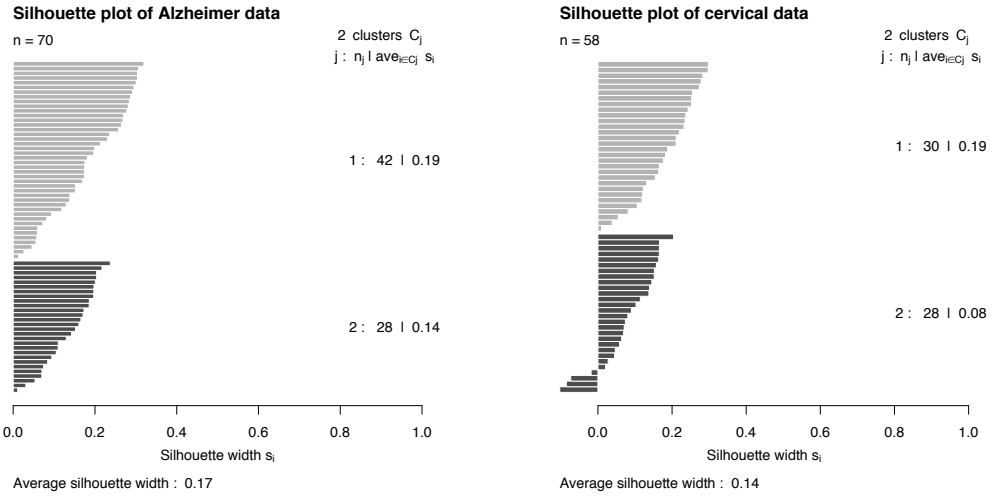


Figure 4.14. Principal components plot for vst transformed values.



(a) **(b)**

Figure 4.15. Silhouette plot for vst transformed values of cervical and alzheimer data.

used to measure how accurate clustering method is while silhouette measures can be used to measure within cluster consistency.

5. DISCUSSION

Microarrays and RNA sequencing are two popular and widely used technologies for gene expression studies. Microarray technology is cost effective while RNA sequencing generates less noisy data. Moreover, RNA sequencing is more recent technology so that statistical approaches which are specifically developed for classification and clustering of RNA sequencing data are limited. In this thesis, we studied several machine learning algorithms for clustering and classification of RNA sequencing data. In bioinformatics, RNA sequencing data is generally used to detect differentially expressed genes between some conditions, i.e diseased versus healthy, subtypes of a cancer etc. Furthermore, gene expression data can be used for early diagnosis of cancer via classification algorithms, or unknown subgroups can be discovered via clustering algorithms. With the recent developments in gene expression studies, discriminating between conditions by using machine learning algorithms became research of interest.

Classification and clustering of RNA sequencing data is not straightforward as in statistics due to distributional and technical limitations of sequencing technology. Raw RNA sequencing data should be prepared for downstream analysis using several normalization and/or transformation techniques. When data set is preprocessed and ready for classification or clustering, it can be analyzed using one of proposed machine learning algorithms. As can be seen, final results depends on several factors such as selected preprocessing method and machine learning algorithm. For this reason, we investigated the effect of several parameters (e.g sample size, differential expression rate, overdispersion, etc) on model accuracies. We selected count based (NBLDA and PLDA) and continuous classifiers (NSC and vNSC) in the analysis. A comprehensive simulation study is conducted to make a detailed comparison between classifiers. Moreover, we used real data sets and compared results with simulation study.

In a real RNA sequencing experiment, the number of features is generally large. For example, an RNA sequencing experiment may consist of 20531 features if all of known human RNAs are sequenced. Simulated data consist of 10000 features and changing number of samples in order to reflect real RNA sequencing experiments. Working with high dimensional data arises several drawbacks. As the number of features and samples increases, the total computation time for analyzing gene expression data also increases dramatically. Machine learning algorithms may be negatively affected from high dimensionality which results in low model accuracies. To overcome this problem, sparse classifiers might be preferred or a feature selection can be performed before training machine learn-

ing algorithms. We used a hybrid strategy to handle this problem by selecting a subset of all features at first and using selected subset with both sparse and non-sparse classifiers. Another drawback of including all features in the model is that an important part of all features will be included in the classifier even if they do not contribute to discrimination. Hence, it is crucial to work with an optimal feature subset contributing to discrimination (or clustering) function. Different approaches can be used to select a subset of all features such as differential expression and maximum variance filtering. Differential expression analysis aims to find features whose expression levels are significantly different between groups. We selected top 500, 1000 and 2000 features by maximum variance filtering. Features are sorted by their variances in descending order and top features are selected. However, maximum variance filtering might ignore the effect of sequencing depth, and deeply sequenced features might be frequently selected. We did not aim to find differentially expressed features and all classifiers worked under the same conditions. Selecting features by differential expression analysis could have increased overall model accuracies; however, we believe that ordering of classifiers by model accuracies might be similar. For a better conclusion, similar simulation study can be performed by considering differential expression analysis in place of maximum variance filtering. We leave this as a further research topic.

Overdispersion is another important parameter which can significantly change accuracy of classifiers. We considered three different levels of overdispersion parameter, e.g low, medium and high. Overdispersion increases as more biological replicates are sequenced. In this case, NBLDA performs better than any other classifier since it takes overdispersion effect into account. When data is overdispersed up to moderate level, a power transformation can be performed to obtain approximately symmetrically distributed gene expression data. Although power transformation is proposed for PLDA, we extended it to NBLDA and observed that model accuracies increased for both PLDA and NBLDA when power transformation is applied.

6. CONCLUSION

In simulation study and real data examples, we studied several classifiers. Among the selected classifiers, NSC and PLDA were sparse models. These classifiers have built-in variable selection criteria which aims to select optimal subset of given features. Although external feature selection criteria can be applied, some of selected features still may not contribute to discriminant function; hence, these features are eliminated within sparse classifiers. This is important when true number of differentially expressed features are very small. The amount of differentially expressed features is controlled by differential expression rate in our simulation study. We set overall differential expression rate as 1%, 5% and 10%. Under the worst scenario, we expect that 100 out of 10000 features to be differentially expressed between conditions. Hence, remaining part of selected features could be removed by sparse classifiers and model performances can be increased. Our simulation results showed that sparse PLDA generally performed better than non-sparse PLDA. We proposed a sparse version of NBLDA in this thesis. Like PLDA, sparse NBLDA outperforms non-sparse NBLDA. However, sparsity of NBLDA was lower than PLDA due to overdispersion effect. NSC, as being continuous classifier, was the most sparse model; however classification accuracy was relatively lower than discrete models.

Real data results agree with the simulation results. Cervical cancer and kidney cancer data sets were highly overdispersed while Alzheimer disease data set was moderately overdispersed. As a result of overdispersion effect, negative binomial models (NBLDA, NBLDA2, sNBLDA and sNBLDA2) performed better comparing to PLDA and continuous classifiers for kidney cancer and cervical cancer data sets. However, discrete and continuous classifiers performed similar for Alzheimer disease data set as a result of moderate overdispersion. The amount of sparsity was generally lower for NSC and vNSC algorithms. In conclusion, simulation and real data results showed that discrete classifiers are better in prediction performance when data set is overdispersed. Continuous classifiers should be preferred for slightly or moderately overdispersed data sets. Finally, NSC algorithm can be preferred when sparsity criteria has higher priority than prediction accuracy.

Clustering can be performed on samples, features or both. Samples are clustered when exploring unknown subgroups (e.g disease subtypes) is of interest while features are clustered when exploring features with similar gene-level activities is of interest. Furthermore, features and samples are clustered simultaneously in order to explore features which are associated with different subgroups. In this

thesis, we focused on the latter condition. We did not conduct a simulation study for clustering part due to two reasons: (i) the number of features were high yielding that dissimilarity matrices were not able to be calculated while features were being clustered, and (ii) clustering analyses were computationally intensive due to the dimension of simulated data. Therefore, clustering is performed only for real data sets using k-means, hierarchical, negative binomial and Poisson clustering methods. Like classification, discrete and continuous clustering algorithms were selected and compared with each other. We found that discrete models and continuous models based on vst transformed values gave similar clustering performances. However, clustering performances may be affected by selected similarity/dissimilarity measures, number of features, number of samples etc. Therefore, we were not able to make a generalization of clustering model performances unless it is supported with a comprehensive simulation results. We leave this as a further research topic.

Over the last twenty years, sequencing platforms evolved rapidly and lots of statistical methods have been proposed for analyzing gene expression data. For example, as sequencing platforms able to produce longer reads, say average read length of 1000 nucleotides, current methodology might lack of power during the analysis. Furthermore, experimental and technical advances require new and novel statistical approaches for differential expression, classification and clustering. According to the results, we conclude that modeling RNA sequencing data with discrete models generally performed better than working with continuous models for transformed raw counts. We proposed a novel extension of NBLDA to sparse model by truncating overdispersion parameter at a given threshold. However, this proposal can be modified by using soft thresholding algorithm for overdispersion and differential expression parameter simultaneously. In clustering part, it is possible to extend Poisson dissimilarity measure to negative binomial case; hence, negative binomial clustering approach can be proposed for clustering RNA sequencing data without a transformation. We leave these two topics as further researches.

BIBLIOGRAPHY

1. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–537.
3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998;95(25):14863–14868.
4. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503–511.
5. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006;7(1):3.
6. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research. 2008;18(9):1509–1517.
7. Hahne F, Huber W, Gentleman R, Falcon S. Bioconductor case studies. New York: Springer; 2008.
8. Witten DM. Classification and clustering of sequencing data using a Poisson model. Annals of Applied Statistics. 2011;5:2493–2518.
9. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. 2004;3(1):1–25.
10. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 2001;98(9):5116–5121.
11. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003;52(1):91–118.
12. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology. 2002;3(7).
13. Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. Bioinformatics. 2003;19(9):1090.
14. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences. 2000;97(1):262–267.
15. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods

for the classification of tumors using gene expression data. Journal of the American Statistical Association. 2002;97(457):77–87.

16. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences. 2002;99(10):6567–6572.

17. Zararsız G. Development and application of machine learning approaches for RNA-seq classification. PhD thesis, Hacettepe University; 2015.

18. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. 2008 7;5(7):613–619.

19. Han X, Wu X, Chung WY, Li T, Nekrutenko A, Altman NS, et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. Proceedings of the National Academy of Sciences. 2009;106(31):12741–12746.

20. Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, et al. A strand-specific RNA-seq analysis of the transcriptome of the Typhoid Bacillus Salmonella Typhi. PLoS Genetics. 2009;5(7):1–13.

21. Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, et al. Conserved developmental transcriptomes in evolutionarily divergent species. Genome Biology. 2010;11(3):R35.

22. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13(3):523.

23. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010;26(1):136.

24. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–1349.

25. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;11(10):R106.

26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550.

27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139.

28. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;15(2):R29.

29. Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Zararsiz GE, Duru IP, et al.

A comprehensive simulation study on classification of RNA-Seq data. PLoS One. 2017;12(8).

30. Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Ozturk A, Karaagaoglu AE. MLSeq: Machine learning interface for RNA-seq data; 2018, R package version 1.20.3.

31. Dong K, Zhao H, Tong T, Wan X. NBLDA: Negative binomial linear discriminant analysis for RNA-seq data. BMC Bioinformatics. 2016;17(1):369.

32. Si Y, Liu P, Li P, Brutnell TP. Model-based clustering for RNA-seq data. Bioinformatics. 2014;30(2):197.

33. Liu P, Si Y. Statistical analysis of Next Generation Sequencing data. Datta S, Nettleton E, editors, Springer; 2015.

34. Reeb PD, Bramardi SJ, Steibel JP. Assessing dissimilarity measures for sample-based hierarchical clustering of RNA sequencing data using Plasmode datasets. PLoS One. 2015;10(7):1–18.

35. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harbor protocols. 2015 04;2015(11):951–969. https://www.ncbi.nlm.nih.gov/pubmed/25870306.

36. Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014 02;30(3):301–304.

37. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010;11(3):R25.

38. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. Science. 2008;321(5891):956–960.

39. 't Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Research. 2008;36(21):e141.

40. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008;5(7):621–628.

41. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 2008;9(2):321–332.

42. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009 Apr;4(1):14.

43. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2012;31:46 EP.

44. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.

BMC Bioinformatics. 2010;11(1):94.

45. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics. 2017;p. bbx008.

46. Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11(1):422.

47. Datta S, Nettleton D, editors. Statistical analysis of next generation sequencing data. Frontiers in probability and the statistical sciences, Springer; 2014.

48. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science. 2003;18(1):104–117.

49. Zararsiz G, Goksuluk D, Klaus B, Korkmaz S, Eldem V, Karabulut E, et al. voomDDA: Discovery of diagnostic biomarkers and classification of RNA-seq data. PeerJ. 2017;5:e3890.

50. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–1401.

51. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature. 2014;509:371 EP. http://dx.doi.org/10.1038/nature13173.

52. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science. 2015;347(6226):1138–1142.

53. Chen Y, McCarty D, Ritchie M, Robinson M, Smith GK. edgeR: differential expression analysis of digital gene expression data. Bioconductor, 3.22.3 ed.; 2018.

54. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. 1971;66(336):846–850.

55. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193–218.

56. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53 – 65. http://www.sciencedirect.com/science/article/pii/0377042787901257.

57. Witten D, Tibshirani R, Gu SG, Fire A, Lui WO. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biology. 2010;8(1):58.

58. Agresti A. Categorical data analysis. Hoboken, NJ: Wiley; 2002.

59. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association. 2002;97(457):77–87.

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

61. Landau WM, Liu P. Dispersion Estimation and Its Effect on Test Performance in RNA-seq Data Analysis: A Simulation-Based Comparison of Methods. PLoS One. 2013 12;8(12):1–16.

62. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second ed. New York: Springer; 2009.

63. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. 1979;74(368):829–836.

64. Oehlert GW. A Note on the Delta Method. The American Statistician. 1992;46(1):27–29.

65. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012;40(10):4288–4297.

66. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010;26(17):2176–2182.

67. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biology. 2013;14(7):R78.

68. Goksuluk D, Zararsiz G, Korkmaz S, Karaagaoglu AE. NBLDA: Negative Binomial Linear Discriminant Analysis; 2018, `https://CRAN.R-project.org/package=NBLDA`, R package version 0.99.0.

69. Saleem M, Padmanabhuni SS, Ngomo ACN, Almeida JS, Decker S, Deus HF. Linked Cancer Genome Atlas Database. In: Proceedings of the 9th International Conference on Semantic Systems I-SEMANTICS '13, New York, NY, USA: ACM; 2013. p. 129–134.

70. Goyal R, Gersbach E, Yang XJ, Rohan SM. Differential Diagnosis of Renal Tumors With Clear Cytoplasm: Clinical Relevance of Renal Tumor Subclassification in the Era of Targeted Therapies and Personalized Medicine. Archives of Pathology & Laboratory Medicine. 2013;137(4):467–480.

APPENDIX

Appendix - 1: Originality Report of the Thesis

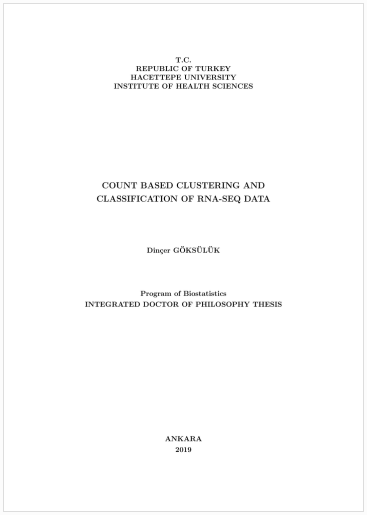
turnitin

Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

Gönderen:	Dinçer Göksülük
Ödev başlığı:	COUNT BASED CLUSTERING AND...
Gönderi Başlığı:	Count based clustering and classif...
Dosya adı:	Tez_Full_Turnitin.pdf
Dosya boyutu:	3.18M
Sayfa sayısı:	72
Kelime sayısı:	22,817
Karakter sayısı:	115,042
Gönderim Tarihi:	05-Şub-2019 05:18PM (UTC+0300)
Gönderim Numarası:	1073355309



T.C.
REPUBLIC OF TURKEY
HACETTEPE UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

COUNT BASED CLUSTERING AND
CLASSIFICATION OF RNA-SEQ DATA

Dinçer GÖKSÜLÜK

Program of Biostatistics
INTEGRATED DOCTOR OF PHILOSOPHY THESIS

ANKARA
2019

Turnitin Orijinallik Raporu

İşleme kondu: 05-Şub-2019 17:23 +03
NUMARA: 1073355309
Kelime Sayısı: 22817
Gönderildi: 1

Count based clustering and classification of RNA-seq data Dinçer Göksülük tarafından

Benzerlik Endeksi	Kaynağa göre Benzerlik	
%7	Internet Sources:	%4
	Yayınlar:	%5
	Öğrenci Ödevleri:	%2

1% match (08-Nis-2018 tarihli internet)
https://peerj.com/articles/3890/

1% match (06-Ara-2018 tarihli internet)
http://www.istkon.net/v2/wp-content/uploads/2017/12/ISTKON10-Abstract-Book-1.pdf

1% match (yayınlar)
Gökmen Zararsız, Dincer Goksuluk, Selcuk Korkmaz, Vahap Eldem, Gozde Erturk Zararsiz, Izzet Parug Duru, Ahmet Ozturk. "A comprehensive simulation study on classification of RNA-Seq data", PLOS ONE, 2017

1% match (yayınlar)
"Statistical Analysis of Next Generation Sequencing Data", Springer Nature America, Inc, 2014

< 1% match (yayınlar)
Law, Charity W, Yunshun Chen, Wei Shi, and Gordon K Smyth. "voom: precision weights unlock linear model analysis tools for RNA-seq read counts", Genome Biology, 2014.

< 1% match (29-Haz-2013 tarihli internet)
http://www.szkolenie.glogow.pl/konin/w/pl/Puchar_%C5%9Awiata_w_skokach_narciarskich

< 1% match (06-Tem-2015 tarihli öğrenci ödevleri)
Submitted to TechKnowledge Turkey on 2015-07-06

< 1% match (04-Kas-2012 tarihli internet)
http://maszyny-ciecie.bialowieza.pl/wiki/w/pl/Puchar_%C5%9Awiata_w_skokach_narciarskich

< 1% match (yayınlar)
"Classification — the Ubiquitous Challenge", Springer Nature America, Inc, 2005

< 1% match (yayınlar)
Gokmen Zararsiz, Dincer Goksuluk, Bernd Klaus, Selcuk Korkmaz, Vahap Eldem, Erdem Karabulut, Ahmet Ozturk. "voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data", PeerJ, 2017

< 1% match (yayınlar)
Vahap Eldem, Gokmen Zararsiz, Tunahan Taşçi, Izzet Parug Duru, Yakup Bakir, Melike Erkan. "Chapter 3 Transcriptome Analysis for Non-Model Organism: Current Status and Best-Practices", InTech, 2017

< 1% match (08-Şub-2014 tarihli internet)
http://www.mtome.com/Publications/JMLR/jmlr-vol7-partB.pdf

< 1% match (20-Eyl-2017 tarihli internet)
http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=4708&context=etd

< 1% match (23-Nis-2011 tarihli internet)
http://www.icm.jhu.edu/theses/theses/Xu.pdf

< 1% match (18-Haz-2003 tarihli internet)
http://www.stata.com/stb/stb7/sbe6/mcc3i.hlp

< 1% match (03-May-2016 tarihli öğrenci ödevleri)
Submitted to National University of Singapore on 2016-05-03

< 1% match (yayınlar)
Advances in Experimental Medicine and Biology, 2015.

< 1% match (yayınlar)
Qingyang Zhang. "Classification of RNA-Seq data via Gaussian copulas", Stat, 2017

< 1% match (yayınlar)
"Classification in BioApps", Springer Nature, 2018

< 1% match (yayınlar)
"Metaheuristic Clustering", Springer Nature America, Inc, 2009

< 1% match (25-Oca-2017 tarihli internet)
http://www.mathworks.com/help/stats/nonlinearmodel.fit.html?nocookie=true&requestedDomain=www.mathworks.com

< 1% match (yayınlar)
"Proceedings of ELM-2015 Volume 1", Springer Nature America, Inc, 2016

< 1% match (07-Nis-2016 tarihli öğrenci ödevleri)
Submitted to Brunel University on 2016-04-07

CURRICULUM VITAE

Dinçer GÖKSÜLÜK

Personal Data

Place and Date of Birth:	Çorum, Turkey | 28 September 1985
Address:	Department of Biostatistics School of Medicine, Hacettepe University Ankara - Turkey
Phone:	+90 312 305 1467
email:	dincer.goksuluk@gmail.com

Work Experience & Positions

Current Sep 2015	Co-founder & Programmer at TURCOSA Analitik, Kayseri - Turkey www.turcosa.com.tr
Current Sep 2012	Research Assistant at Hacettepe University, Ankara - Turkey *Department of Biostatistics*
Current May 2017	Publication Officer *International Biometric Society - Eastern Mediterranean Region (IBS - EMR)*

Education

Current	Combined PhD in Biostatistics, **Hacettepe University**, Ankara - Turkey Thesis: "Count based clustering and classification of RNA-seq data" Advisor: Prof. A. Ergun Karaagaoglu
August 2011	Master of Science in Statistics, **Dokuz Eylul University**, Izmir - Turkey Thesis: "Penalized logistic regression" Advisor: Prof. Aylin Alin
June 2008	Undergraduate Degree in Statistics, **Dokuz Eylul University**, Izmir - Turkey Gpa: 80.33/100

Research Areas

In Statistics
- Machine Learning
- Multivariate Statistics
- General Linear Models (GLMs)
- Design of Experiments
- Longitudinal Data Analysis

In Bioinformatics
- Next Generation Seguencing (NGS)
- Transcriptomics
- Proteomics
- Drug Discovery (Homology Modelling)
- Microarrays

Languages

Turkish:	Native
English:	Fluent

Computer Skills

Beginner:	Java Scripts, html, SAS, STATA,
Intermediate:	Matlab, Mapple, Orange
Upper-intermediate:	Linux, LaTeX, Excel, Word, PowerPoint
Advanced:	R, SPSS, MINITAB

Courses (Taken)

February 2015 Sequential Equation Modelling
Dept. of Biostatistics, **Osmangazi University**, Eskisehir - Turkey
Lecturer: Prof. Kazım Özdamar

May 2014 Generalized Linear Mixed Models with Applications in Medicine
Dept. of Biostatistics, **Ege University**, Izmir - Turkey
Cooperation with University of Southampton and University of Reading
Lecturers: Prof. Dankmar Böhning
Dr. Stefanie Biedermann
Mr. James Gallagher

January 2014 Evidence-Based Medicine for Clinicians
Weill Cornell Medical Collage

Courses (Given)

June 2018 Biostatistics Course
Hacettepe University, Ankara - Turkey

April 2018 Biostatistics Course - II
Çocuk Nefroloji Derneği, Ankara - Turkey

April 2018 Biostatistics Course
Hacettepe University, Ankara - Turkey

October 2017 Applied predictive modeling with R
19^{th} *National and* 2^{nd} *International Biostatistics Congress*, Antalya - Turkey

March 2017 Biostatistics Course - I
Çocuk Nefroloji Derneği, Ankara - Turkey

October 2016 Bioinformatics analysis with R
18^{th} *National Biostatistics Congress*, Antalya - Turkey

Awards & Honours

2018 Sıfır yığılımlı (zero-inflated) kesikli verilerin modellenmesinde regresyon yaklaşımları
International Biostatistics Congress - Selected Oral Presentation, 1^{st} place
SANKO University & Biostatistics Society, Gaziantep - Turkey

2017 Machine learning based virtual screening in drug discovery
International Biostatistics Congress - Selected Oral Presentation, 1^{st} place
Biostatistics Society, Antalya - Turkey

2017 MLSeq 2.0: Machine learning interface for RNA-Sequencing data
International Biostatistics Congress - Selected Oral Presentation, 2^{nd} place
Biostatistics Society, Antalya - Turkey

2017 2×2 tablolar için kapsamlı bir web yazılımı
International Biostatistics Congress - Selected Short Oral Presentation, 2^{nd} place
Biostatistics Society, Antalya - Turkey

2016 compSurv: An interactive web-tool for survival analysis
National Biostatistics Congress - Selected Oral Presentation, 3^{rd} place
Biostatistics Society, Antalya - Turkey

2015 Decision support system for differential diagnosis of nontraumatic acute abdomen
Gevher Nesibe Research Prize - Best Oral Presentation
Erciyes University, Kayseri - Turkey

2015 Diagonal discriminant analysis for gene-expression based tumor classification
Best Oral Presentation
3^{rd} **International Conference on Bioinformatics and Computational Biology**, Hong Kong

2014 Classification of RNA-Seq data via bagging support vector machines
Young Statistician Showcase - Best Oral Presentation
27^{th} **International Biometric Conference (IBC)**, Florence - Italy
Fund: 3000$

2014 Data mining for Next Generation Sequencing data analysis
Gevher Nesibe Research Prize - Best Oral Presentation
Erciyes University, Kayseri - Turkey

2014 A novel approach for the classification of RNA-Seq based gene expression data. MLSeq: an R/Bioconductor package
Best Oral Presentation
16^{th} **National Biostatistics Symposium**, Antalya - Turkey

Projects

1. Kullanıcı dostu, kişiselleştirilmiş ve bulut tabanlı istatistiksel analiz sistemi
   Project Category: TUBITAK-1512
   Coordinator: Assist. Prof. Selçuk Korkmaz
   Position: Researcher/Partner
   Supporter: The Scientific and Technological Research Council of Turkey (TUBITAK)
2. Kullanıcı dostu, kişiselleştirilmiş ve bulut tabanlı istatistiksel analiz sistemi
   Project Category: TUBITAK-1005
   Coordinator: Assist. Prof. Gökmen Zararsiz
   Position: Researcher/Partner
   Supporter: The Scientific and Technological Research Council of Turkey (TUBITAK)
3. Novel Statistical Learning Algorithms for the Classification of RNA-Seq Data
   Coordinator: Prof. Ahmet Öztürk
   Position: Researcher
   Supporter: Scientific Research Projects Coordination Unit - Erciyes University, Turkey
   Fund: $\sim$ 11,195 $

Researches (SCI, SCI-Exp, SSCI, AHCI)

1. Bal C, Öztürk A, Çiçek B, Özdemir A, Zararsız G, Ünalan D, et al. The Relationship Between Blood Pressure and Sleep Duration in Turkish Children: A Cross-Sectional Study. Journal of Clinical Research in Pediatric Endocrinology. 2018 03;10(1):51–58.
2. Karaismailoglu E, Konar NM, **Goksuluk D**, Karaagaoglu AE. Factors effecting the model performance measures area under the ROC curve, net reclassification improvement and integrated discrimination improvement. Communications in Statistics - Simulation and Computation. 2018;0(0):1–13. https://doi.org/10.1080/03610918.2018.1458135.
3. Korkmaz S, Duarte JM, Prlić A, **Goksuluk D**, Zararsiz G, Saracbasi O, et al. Investigation of protein quaternary structure via stoichiometry and symmetry information. PLoS One. 2018 06;13(6):1–20.
4. Zararsiz G, **Goksuluk D**, Klaus B, Korkmaz S, Eldem V, Karabulut E, et al. voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. PeerJ. 2017;5:e3890.

5. Korkmaz S, **Goksuluk D**, Zararsiz G, Karahan S. geneSurv: An interactive web-based tool for survival analysis in genomics research. Computers in Biology and Medicine. 2017;89(Supplement C):487 – 496. http://www.sciencedirect.com/science/article/pii/S001048251730286X.

6. Zararsiz G, **Goksuluk D**, Korkmaz S, Eldem V, Zararsiz GE, Duru IP, et al. A comprehensive simulation study on classification of RNA-Seq data. PLoS One. 2017 08;12(8):1–19. https://doi.org/10.1371/journal.pone.0182507.

7. Firat A, Alemdaroglu KB, Ozmeric A, Yucens M, **Goksuluk D**. Morphometric study of the true S1 and S2 of the normal and dysmorphic sacralized sacra. Turkish Journal of Medical Sciences. 2017;47(3):954–959.

8. Dogru HB, Avcu N, Akkaya N, Yilanci HO, **Goksuluk D**. Applicability of Cameriere's and Drusini's age estimation methods to a sample of Turkish adults. Dentomaxillofacial Radiology. 2017;46(7):20170026. https://doi.org/10.1259/dmfr.20170026, pMID: 28707524.

9. Yilanci HO, Akkaya N, **Goksuluk D**. A preliminary study of dental patterns in panoramic radiography for forensic identi cation. Romanian Journal of Legal Medicine. 2017;25:75–81.

10. Yildirim TT, Guncu GN, **Goksuluk D**, Tozum MD, Colak M, Tozum TF. The effect of demographic and disease variables on Schneiderian membrane thickness and appearance. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. 2017;124(6):568–576. http://dx.doi.org/10.1016/j.oooo.2017.09.002.

11. Unal E, Kisacik P, Arin G, Karabulut E, **Goksuluk D**, Vardar NY, et al. AB1198-HPR presentation of a new scale assessing the biopsychosocial aspects of healing properties in rheumatic patients. Annals of the Rheumatic Diseases. 2017;76(Suppl 2):1531–1531. http://ard.bmj.com/content/76/Suppl_2/1531.2.

12. Bal C, Ozturk A, Cicek B, Ozdemir A, Zararsiz G, Unalan D, et al. The relationship between blood pressure and sleep duration in Turkish children: A cross-sectional study. Journal of Clinical Research in Pediatric Endocrinology. 2017 June;https://doi.org/10.4274/jcrpe.4557.

13. **Goksuluk D**, Korkmaz S, Zararsiz G, Karaagaoglu AE. easyROC: An interactive web-tool for ROC curve analysis using R language environment. The R Journal. 2016;8(2):213–230. https://journal.r-project.org/archive/2016/RJ-2016-042/index.html.

14. Zararsiz G, Akyildiz HY, **Goksuluk D**, Korkmaz S, Ozturk A. Statistical learning approaches in diagnosing patients with nontraumatic acute abdomen. Turkish Journal of Electrical Engineering and Computer Sciences. 2016;24(5):3685–3697.

15. Akkaya N, Yilanci HO, **Goksuluk D**. Applicability of Demirjian's four methods and Willems method for age estimation in a sample of Turkish children. Legal Medicine. 2015;17(5):355–359.

16. Korkmaz S, Zararsiz G, **Goksuluk D**. MLViS: A web tool for machine learning-based virtual screening in early-phase of drug discovery and development. PLoS One. 2015;10(4):e0124600.

17. Korkmaz S, Zararsiz G, **Goksuluk D**. Drug/nondrug classification using Support Vector Machines with various feature selection strategies. Computer Methods and Programs in Biomedicine. 2014;117(2):51–60.

18. Korkmaz S, **Goksuluk D**, Zararsiz G. MVN: An R package for assessing multivariate normality. The R Journal. 2014;6(2):151–162. https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf.

Researches (Other)

1. Zararsiz G, Zararsiz GE, Ozturk A, **Goksuluk D**, Korkmaz S, Eldem V, et al. Genome wide gene expression profiling and molecular classification of renal cell cancer subtypes. Journal of Advances in Information Technology. 2017;8(1):10–16.

2. Unal E, Arin G, Karaca NB, Kiraz S, Akdogan A, Kalyoncu U, et al. Romatizmalı hastalar için bir yaşam kalitesi ölçeğinin geliştirilmesi: madde havuzunun oluşturulması. Journal of Exercise Therapy and Rehabilitation. 2017;4(2):67–75.

3. Tanacan A, Aydin E, Cakar AN, Cagan M, **Goksuluk D**, Beksac MS. The effect of prenatal invasive tests on neonatal birthweight. Gynecology Obstetrics & Reproductive Medicine. 2016;21(3).

4. Zararsiz G, Korkmaz S, **Goksuluk D**, Eldem V, Ozturk A. Diagonal discriminant analysis for gene-expression based tumor classification. Journal of Advances in Information Technology. 2015;6(2):59–62.

5. Korkmaz S, **Goksuluk D**, Zararsiz G. MVN: an R package for assessing multivariate normality. CRAN, package version 40. 2014;http://CRAN.R-project.org/package=MVN.

6. Zararsiz G, **Goksuluk D**, Korkmaz S, Eldem V, Duru IP, Ozturk A. MLSeq: Machine learning interface for RNA-Seq data. Bioconductor, package version 16. 2014;http://www.bioconductor.org/packages/devel/bioc/vignettes/MLSeq/inst/doc/MLSeq.pdf.

Conference Proceedings (International)

1. Goksuluk MB, **Goksuluk D**, Karaagaoglu AE (2018). "Sıfır yığılımlı (zero-inflated) kesikli verilerin modellenmesinde regresyon yaklaşımları". *The* 20^{th} *National and* 3^{rd} *International Biostatistics Congress*, October 26–29, Gaziantep - Turkey.

2. **Goksuluk D**, Basol M, Haklı DA (2017). "sNBLDA: Sparse negative binomial linear discriminant analysis ". *The* 10^{th} *International Statistics Congress*, December 6–8, Ankara - Turkey.

3. Basol M, **Goksuluk D**, Karaagaoglu AE (2017). "Bootstrap güven aralığı yöntemlerinin karşılaştırılması". In Turkish. *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

4. Zararsiz G, **Goksuluk D**, Korkmaz S, Eldem V, Klaus B, Ozturk A (2017). "MLSeq 2.0: Machine learning interface for RNA sequencing data". *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

5. Korkmaz S, Zararsiz G, **Goksuluk D**, Sut N (2017). "Machine learning based virtual screening in drug discovery". *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

6. Cavusoglu M, Zararsiz G, Zararsiz GE, Korkmaz S, **Goksuluk D**, Mazicioglu M, Ozturk A (2017). "GAMLSS modellerinin model yeterliliğinin belirlenmesinde Worm eğrileri". In Turkish. *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

7. Eldem V, Zararsiz G, Korkmaz S, **Goksuluk D**, Zararsiz GE, Bilgin H, Ozturk A (2017). "A Glimse at long-read sequencing and megabase-sized scaffolding approaches". *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

8. Bilgin H, Zararsiz G, Ozkaya V, Cicek B, Zararsiz GE, Korkmaz S, **Goksuluk D**, Ozturk A (2017). "Improved estimation of body fat percentage via machine-learning approaches". *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

9. Unlusavuran M, Zararsiz G, Ipekten F, Zararsiz GE, Korkmaz S, **Goksuluk D**, Dogan HO, Eldem V, Ozturk A (2017). "Metabolomik verilerin sınıflandırılmasında kısmi en küçük kareler ayırma analizi yaklaşımı". In Turkish. *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

10. Ipekten F, Zararsiz G, Unlusavuran M, Zararsiz GE, Korkmaz S, **Goksuluk D**, Dogan HO, Eldem V, Ozturk A (2017). "Metabolomik biyobelirteçlerinin tespitinde ANOVA-PCA yaklaşımı". In Turkish. *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

11. Durmuscelebi A, Zararsiz G, Korkmaz S, Bilgin H, **Goksuluk D**, Zararsiz GE, Elmali F, Ozturk A (2017). "2×2 tablolar için kapsamlı bir web yazılımı (Poster & Short talk)". In Turkish. *The* 19^{th} *National and* 2^{nd} *International Biostatistics Congress*, October 25–28, Antalya - Turkey.

12. Ozturk A, Cicek B, Zararsiz G, Zararsiz GE, Korkmaz S, **Goksuluk D**, Mazicioglu M (2017). "Association of internet addiction and depression in graduate students". *The* 5^{th} *Conference of the Association of General Practice / Family Medicine of South-East Europe*, May 25–28, Budva - Montenegro.

13. Korkmaz S, Ozturk A, Cicek B, Erceyes DU, Zararsiz G, Zararsiz GE, **Goksuluk D**, Mazicioglu M (2017). "The relationship between blood pressure and sleep duration in Turkish children: a cross-sectional study". *The* 5^{th} *Conference of the Association of General Practice / Family Medicine of South-East Europe*, May 25–28, Budva - Montenegro.

14. Ozturk A, Cicek B, Zararsiz G, Korkmaz S, **Goksuluk D**, Mazicioglu M (2017). "Wrist circumference and frame size percentiles in 6–17 year old Turkish children and adolescents in Kayseri (Poster)". *The* 5^{th} *Conference of the Association of General Practice / Family Medicine of South-East Europe*, May 25–28, Budva - Montenegro.

15. Basol M, **Goksuluk D**, Karaagaoglu AE (2017). "A comprehensive simulation study for comparison of statistical methods in MRMC ROC studies". *The* 9^{th} *Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS)*, May 8-12, Thesseloniki - GREECE.

16. Hakli DA, **Goksuluk D**, Karabulut E (2017). "How to overcome class imbalance problem (Poster)". *The* 9^{th} *Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS)*, May 8-12, Thesseloniki - GREECE.

17. Dogru HB, Stelt PF, Kamburoglu K, **Goksuluk D**, Avcu N (2017). "Assessment of the frontal sinuses using CBCT for gender determinat ion in samples of Dutch and Turkish individuals". *The* 21^{st} *International Congress of Dental and Maxillo-Facial Radiology*, April 26-29, Kaohsiung - TAIWAN.

18. **Goksuluk D**, Korkmaz S, Zararsiz G (2015). "easyROC: an interactive web-tool for ROC analysis". *The useR! Conference 2015*, June 30 - July 3, Aalborg - DENMARK.

19. Zararsiz G, Akyildiz HY, **Goksuluk D**, Korkmaz S, Ozturk A (2015). "DDNAA: Decision support system for differential diagnosis of nontraumatic acute abdomen". *The useR! Conference 2015*, June 30 - July 3, Aalborg - DENMARK.

20. Zararsiz G, **Goksuluk D**, Korkmaz S, Eldem V, Duru IP, Unver T, Ozturk A (2015) "MLSeq: Machine learning interface for RNA-Seq data". *The useR! Conference 2015*, June 30 - July 3, Aalborg - DENMARK.

21. **Goksuluk D**, Zararsiz G, Korkmaz S, Karaagaoglu AE (2015). "Negative binomial linear discriminant analysis for the classification of RNA-Seq data". *The* 8^{th} *Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS)*, 11-15 May, Cappadocia - TURKEY.

22. Zararsiz G., Akyildiz HY., **Goksuluk D.**, Korkmaz S., Ozturk A. (2015). "Statistical learning approaches in diagnosing patients with nontraumatic acute abdomen". *The* 8^{th} *Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS)*, 11-15 May, Cappadocia - TURKEY.

23. Konar NM, Karaismailoglu E, **Goksuluk D**, Karaagaoglu AE (2015). "The effect of correlation structure between diagnostic tests on net reclassification improvement (NRI) and integrated discrimination improvement (IDI)". *The* 8^{th} *Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS)*, 11-15 May, Cappadocia - TURKEY.

24. Zararsiz G, Eldem V, Korkmaz S, **Goksuluk D**, Duru IP, Unver T, Karabulut E, Ozturk A (2015). "Diagonal discriminant analysis for gene-expression based tumor classification". *International Conference on Bioinformatics and Computational Biology (ICBCB 2015)*, 12-13 March, Hong Kong, HONGKONG.

25. **Goksuluk D**, Korkmaz S, Zararsiz G, Karaagaoglu AE (2014). "Ensemble machine learning approach for biomarker discovery using Mass-Spectrometry based proteomics data". 27^{th} *International Biometric Conference (IBC)*, 6-11 July, Florence - ITALY.

26. Zararsiz G, **Goksuluk D**, Korkmaz S, Eldem V, Unver T (2014). "Classification of RNA-Seq data via bagging support vector machines". 27^{th} *International Biometric Conference (IBC)*, 6-11 July, Florence - ITALY.

27. Korkmaz S, Zararsiz G, **Goksuluk D**, Saracbasi O (2014). "Homology modeling using machine learning approach". 27^{th} *International Biometric Conference (IBC)*, 6-11 July, Florence - ITALY.

28. Korkmaz S, **Goksuluk D**, Zararsiz G (2013). "bbRVM: An R package for ensemble classification approaches of Relevance Vector Machine". *The R User Conference*, 10-12 July, Albacete - SPAIN.

29. **Goksuluk D**, Gurler S, Eliiyi DT (2013). "A Bayesian approach to preventive maintenance optimization in unreliable systems". 26^{th} *European Conference on Operational Research (EURO-INFORMS)*, 1-4 July, Rome - ITALY.

30. **Goksuluk D**, Alin A (2011). "Penalized logistic regression when multicollinearity exists among the predictors". *New Developments in Theory and Applications of Statistics (NEDETAS)*, Middle East Technical University, Ankara - TURKEY.

Conference Proceedings (National)

1. **Goksuluk D**, Korkmaz S, Zararsiz G and Karahan S (2016). "compSurv: An interactive web-tool for survival analysis". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

2. Basol M, **Goksuluk D**, Karaagaoglu AE (2016). "Extensions to conventional ROC analysis: FROC and AFROC". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

3. Korkmaz S, Rose PW, Duarte JM, Prlic A, **Goksuluk D**, Zararsiz G., Saracbasi O (2016). "Determining the incorrect protein structures and its ratio in PDB database". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

4. Zararsiz G, Korkmaz S, Eldem V, **Goksuluk D**, Zararsiz GE, Unlusavuran M, Ozturk A (2016). Independent hypothesis weighting in differential expression analysis of RNA-Seq data. 18^{th} *National Biostatistics Congress*, 26-29 October 2016, Antalya, TURKEY.

5. Ozturk A, Zararsiz G, Cicek B, Mazicioglu MM, Zararsiz GE, Unlusavuran M, **Goksuluk D**, Korkmaz S, Kurtoglu S (2016). "Using Hattori graph to assess the body measures of children aged 6 to 17 in province Kayseri". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

6. Durmuscelebi A, Zararsiz G, **Goksuluk D**, Korkmaz S, Ozturk A (2016). "DataOrganizer: A web tool for data manupilation". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

7. Hakli DA, Karabulut E, **Goksukuk D** (2016). "Algorithms to overcome class imbalance problem". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

8. Unlusavuran M, Zararsiz G, Zararsiz GE, **Goksuluk D**, Korkmaz S, Kavcu BO, Ozturk A (2016). "Unsupervised Random Forest algorithm for clustering metabolomics data". 18^{th} *National Biostatistics Congress*, 26-29 October, Antalya, TURKEY.

9. **Goksuluk D**, Korkmaz S, Zararsiz G, Karaagaoglu AE (2014). "Ensemble machine learning approach for biomarker discovery using Mass-Spectrometry based proteomics data". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

10. Zararsiz G, **Goksuluk D**, Korkmaz S, Eldem V, Duru IP, Unver T, Ozturk A (2014). "A new approach for the classification of RNA-Seq based gene expression data, MLSeq: an R/Bioconductor package". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

11. Korkmaz S, Zararsiz G, **Goksuluk D**, Saracbasi O (2014). "Detecting the homolog proteins via machine learning algorithms". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

12. Korkmaz S, **Goksuluk D**, Zararsiz G (2014). "MVN: An R packge for assessing the multivariate normality". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

13. **Goksuluk D**, Zararsiz G, Korkmaz S, Eldem V, Duru IP (2014). "Comparing the computation times of significance tests used for RNA-Seq data (Poster)". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

14. Zararsiz G, Eldem V, Korkmaz S, **Goksuluk D**, Duru IP, Unver T, Ozturk A (2014). "De Novo gene sequencing analysis using De-Brujin plots (Poster)". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

15. Zararsiz G, Korkmaz S, **Goksuluk D**, Eldem V, Ozturk A (2014). "Statistical evaluation of RNA-Seq data (Poster)". 16^{th} *National Biostatistics Congress*, Ankara University, 10-12 September, Antalya - TURKEY.

16. Gurler S, Eliiyi DT, **Goksuluk D**, Sahin A (2010). "Determining preventive maintenance times in unreliable systems with Bayesian method". *Istatistik Günleri Sempozyumu (Statistics Days Symposium)*, Middle East Technical University, Ankara - TURKEY.

17. **Goksuluk D**, Alin A, Emrem E, Telci BK (2008). "Two alternative regression methods when multicollinearity exists: Principal Component Regression (PCR) and Partial least Squares Regression (PLS)". 5^{th} *Statistics Colloquium*, Selçuk University, 6-7 May, Konya - TURKEY.

Attanded Congresses & Conferences (International)

1. The 10^{th} International Statistics Congress, 6–8 December 2017, Ankara, Turkey.
2. The 19^{th} National and 2^{nd} International Biostatistics Congress, 25-28 October 2017, Antalya, Turkey.
3. The 9^{th} Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS), 8-12 May 2017, Thesseloniki, Greece.
4. The R useR Conference, 30 June – 3 July 2015, Aalborg, Denmark.
5. The 8^{th} Conference of Eastern Mediterranean Region International Biometric Society (EMR-IBS), 11-15 May 2015, Cappadocia, Nevsehir, Turkey.
6. 27^{th} International Biometric Conference (IBC), 6-11 July 2014. *International Biometric Society*, Florence, Italy.
7. 26^{th} European Conference on Operational Research (EURO-INFORMS), 1-4 July 2013. *Sapienza University of Rome*, Rome, Italy.
8. The R useR Conference 2013, 10-12 July. *University of Castilla La-Mancha*, Albacete, Spain.
9. New Developments in Theory and Applications of Statistics (NEDETAS). 2-4 May 2011, *Middle East Technical University*, Ankara, Turkey.

Attanded Congresses & Conferences (National)

1. 18^{th} National Biostatistics Congress, 26-29 October 2016, Antalya, Turkey.
2. 16^{th} National Biostatistics Congress, 10-12 September 2014, *Ankara University*, Antalya, Turkey.
3. 15^{th} National Biostatistics Congress, 20-23 August 2013, *Adnan Menderes University*, Aydin, Turkey.
4. VII. Istatistik Gunleri Sempozyumu (IGS), 28-30 June 2010, *Middle East Technical University*, Ankara, Turkey.
5. V. Istatistik Kolokyumu, 6-7 May 2008, *Selcuk University*, Konya, Turkey.

Book Chapters

1. **Goksuluk, D.** Epidemiolojik raporların yorumlanması (Ek-1). pages: 1413-1418. (Gunalp GS. Klinik Jinekolojik Endikronoloji ve Infertilite, 8. Baskı. Güneş Tıp Kitabevi. (Turkish Edition of: Fritz M. A., Speroff L. Clinical Gynecologic Endocrinology and Infertility, Eighth Edition.))

Softwares & Web-tools

1. easyROC: An interactive web-tool for ROC analysis. http://www.biosoft.hacettepe.edu.tr/easyROC
2. DDNAA: Decision support system for differential diagnosis of nontraumatic acute abdomen, http://www.biosoft.hacettepe.edu.tr/DDNAA
3. MVN: a web-tool for assessing multivariate normality. http://www.biosoft.hacettepe.edu.tr/MVN
4. MVN: Multivariate normality tests. (R Package). http://cran.r-project.org/web/packages/MVN/index.html
5. MLViS: Machine-learning based visual screening. http://www.biosoft.hacettepe.edu.tr/MLViS/
6. MLSeq: Machine-learning interface for RNA-Seq data. R/Bioconductor Paketi (R Package). http://www.bioconductor.org/packages/release/bioc/html/MLSeq.html
7. voomDDA: Discovery of diagnostic biomarkers and classification of RNA-Seq data. http://www.biosoft.hacettepe.edu.tr/voomDDA