Charles University in Prague

Faculty of Mathematics and Physics

DOCTORAL THESIS

Bobosharif K. Shokirov

Normality test of the gene expression data

Department of Probability and Mathematical Statistics

Supervisor of the doctoral thesis:	prof. Lev B. Klebanov, DrSc.
Study program:	Mathematics
Specialization:	Probability and Mathematical Statistics

Prague 2015

Acknowledgements

First and foremost I would like to express my deep appreciation to my supervisor Professor Lev B. Klebanov, DrSc., for leading me into this research area, for every piece of advice he provided, for his tremendous support.

I would like to express my appreciation for everything I learned from him during my PhD studies at the Department of Probability and Mathematics Statistics of the Charles University in Prague, for his patience. Although it was not easy, but I have to acknowledge his precision, especially when it comes to the proof of a statement.

Next, I would like to thank the following people from the Department, who in some way supported me to complete my degree.

Professor RNDr. Marie Hušková, DrSc., for her weekly informative “schůzka” on Wednesdays before the seminar of “Asymptotické metody matematické statistiky” starts, for all her advice and support during these long years I spent in the Department of Probability and Mathematical Statistics of MFF UK.

Professor RNDr. Jana Jurečková, DrSc., for her seminar-course “Asymptotické metody matematické statistiky”, for acquainting me with such great sources of mathematical statistics, one of which is, ”Characterizations problems of Mathematical Statistics” by Kagan, A. M., Linnik, Yu. V. and Rao, C. R., and for one year employing me in “Jaroslav Hájek Center for Theoretical and Applied Statistics”.

Prof. RNDr. Jaromír Antoch, CSc., for all his support during this period, starting from the course of “Simulační metody matematické statistiky”, for his patience in repeating each lecture material twice, the second time in the understantable (to me) Russian language, and ending with a tour outside of Prague.

Dr. rer. nat. Jan Kalina, PhD., and Ing. Marek Omelka, PhD., my former colleagues from “Jaroslav Hájek Center for Theoretical and Applied Statistics” for all assistance they provided me, even helping with LaTeX.

Outside of my school, I would like to thank Carol Reed, my mother-in-law and David Goshert, my father-in-law, who despite all the difficulties they had, could find an opportunity to support me for a long period, and for their constant prayers for me to be able to complete my thesis.

Last, but not least, I would like to thank my wonderful wife Lori, the best lady I ever met, for her everyday support and encouragement and my lovely daughter Arvaneh, a small sweet princess, for her sound night sleeps that gave me an opportunity to work, without which this work would never have been completed.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague, 07.06.2015 Bobosharif K. Shokirov

Název práce: Testovani normalnosti dat genovich ekpresse

Autor: Bobosharif K. Shokirov

Katedra: Katedra pravdepodobnosti a matematicka statistika

Vedoucí disertační práce: profesor Lev B. Klebanov, DrSc., KPMS MFF UK

Abstrakt: Tato práce se zabývá testování normality dat genové exprese. Na základě charakterizacni věty normální rozdělení test normalnosti je nahrazen testem sférické stejnoměrnosti. Vzhledem k silné korelace mezi dat genové exprese, test normalnosty se provádí aplikací δ sekvencí. Je dokazano nová charakterizacni věta normálního rozdělení, a na základě toho, test normalnosti se provádí pouzitim Kolmogorovuv test. Získané charakterizacni výsledky pro normální rozdělení jsou rozšířeny do kompletneho typu rozdělení, a na zaklade toho testováno, zda že rozdělení dvou datových souborů genové exprese patří do stejného typu.

Klíčová slova: genové exprese, rekonstrukce rozdělení, maximální invariantni statisticka, sfericky stejnoměrnost, test normálnosti.

Title: Normality test of the gene expression data

Author: Bobosharif K. Shokirov

Department: Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague

Supervisor: Lev B. Klebanov, Professor, DrSc., Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague

Abstract: This thesis deals with a test of normality of gene expressions data. Based on characterization theorems of the normal distribution, the test of normality is replaced by a test of spherical uniformity. Due to strong correlations between the gene expression data, the normality test is conducted with δ sequences. A new characterization theorem of the normal distribution is proven. Based on that, the normality test is conducted using Kolmogorov's test statistic. The obtained characterization results for the normal distribution are extended to the complete type of distributions and based on that, a test is conducted to verify whether the distributions of the two data sets of the gene expressions belong to the same type.

Keywords: gene expression, reconstruction of distribution, maximal invariant statistic, spherical uniformity, normality test.

Contents

Acknowledgements **i**

Contents **iv**

List of Figures **vi**

List of Tables **vii**

Abbreviations **viii**

Introduction **1**

1 Formulation of the problem **5**

5

1.1 Introduction 5

1.2 Does the logarithm of the gene expressions follow the normal law? 6

1.3 A general overview of the logarithm of the gene expression data 7

1.3.1 Biological background 8

1.3.2 Microarray Technology 11

1.4 Dependence between gene expressions 17

1.5 The idea of using many genes 19

2 Reconstruction type of distributions **23**

23

2.1 Introduction 23

2.2 Reconstruction the type of distributions 23

2.2.1 Reconstruction of the additive type 24

2.2.2 Reconstruction of the multiplicative type 25

2.2.3 Reconstruction of the complete type 26

2.3 Characterization of the normal density 26

2.3.1 Characterization theorem by A. A. Zinger 26

2.3.2 Characterization theorem by T. Sakata 27

3 Test of spherical uniformity **32**

32
3.1 Introduction 32
3.2 Statistics for uniformity test on the sphere 32
3.2.1 $\mathfrak{N}$-distances 32
3.2.2 Statistics based on $\mathfrak{N}$-distances 34
3.3 Application of the $\mathfrak{N}$-distance based test to gene expression data 35
3.3.1 Calculating statistic T_n and its p values 35
3.3.2 Uniformity test of the gene expression data on the sphere 36
3.3.3 Modifications of the uniformity test on the sphere 37

4 One-dimensional test and the distribution of its statistic 41

41
4.1 Introduction 41
4.2 The U test 41
4.2.1 The Meijer G-function 42
4.2.2 A characterization theorem of the normal distribution 43
4.2.3 A characterization theorem of the complete type 49
4.3 Application of the Kolmogorov and Kolmogorov - Smirnov's statistics to the gene expression data 51
4.3.1 Results of the normality test 52
4.3.2 Test results of whether the two samples belong to the same type . 53

Conclusion 55

The list of author's publications 58

Bibliography 59

Index 66

List of Figures

1.1 DNA and its building blocks 10

1.2 Histogram of the correlation coefficients 21

3.1 Observations on the unit sphere 40

4.1 Distribution function and the characteristic function of the statistic U . . 49

4.2 Probability density and the characteristic function of the statistic U . . . 49

List of Tables

1.1 Some statistics of the human genome 22

3.1 Results of spherical uniformity test 37
3.2 The results of uniformity test by Algorithm 1 39

4.1 Results of normality test 52
4.2 The test results of whether the two samples belong to the same type . . . 54

Abbreviations

DNA	Deoxyribonucleic acid
cDNA	complemantary DNA
RNA	Ribonucleic acid
mRNA	messenger RNA
tRNA	transcription RNA
i.i.d.	independent and identically distributed
r.v.	random variable
ch.f.	characteristic function
d.f.	distribution function
PCR	Polymerase Chain Reaction
RT-PCR	Reverse Transcription PCR
qRT-PCR	Real-Time Quantitative Reverse Transcription
SAGE	serial analysis of gene expression

Introduction

The discovery of the structure of the deoxyribonucleic acid (DNA) molecule by Watson and Crick (1953) raised the understanding of the molecular mechanisms of genetics to a new level. A wide range of research areas, from medicine to agriculture, received great benefits from it. And the recent new advances of high-throughput microarray technologies changed research in molecular biology in many aspects. The new technologies became a promising field for many research areas, in particular, for those connected with human health.

By the extent of availability data of the gene expressions measurements, data of whole-genome sequence for many mammal species, and especially for humans, the field of biology has become data-driven than ever before. In this meaning, biological and pharmaceutical researchers have been facing a new challenge in the analysis of the new wealth of data. If a biology expert can manage to analyze the results of a small-scale experiment, while the analysis of the results of a large-scale experiment, such as a microarray experiment, requires some other tools and specific skills beyond the biologist's expertise. Basically, this is a starting point from which a collaboration of a biologist and experts from other research areas, in particular from statistics, begins.

One important aspect in genomics is the understanding of the transcriptome, the summation of all RNA transcripts within a given cell. Since the genome of a given organism is a catalog of encoded molecular information, then how that information would be used is of more importance. An individual gene within a genome can be expressed via transcription at different rates, which depending on the stimuli of the cellular system can be up-regulated or down-regulated. Now, capability of quantifying gene expression levels and comparing the abundance of a particular mRNA transcripts make it possible to examine the genome functionality in its entirety.

Quantification of the gene expressions is one of the growing processes in the study of genomics, which starting from low-throughput methods of measurements, such as quantitative northern, southern blots, western blot (or sometimes called immunoblot), Polymerase Chain Reaction (PCR) at the beginning, has grown to the high-throughput methods, such as serial analysis of gene expression (SAGE) and array based approaches, in the present days. The study of the microarray experiment itself is one of the research areas of great interest. Since the publication by Schena (Schena, 1995) interest in this area has been growing at a higher rate. For example, in a study designed by Golub et al. (1999), patients with leukemia were classified into two known subgroups by using just gene expressions. Sørlie et al. (2001) identified five patterns of the gene expression levels in breast cancer and identified them as corresponding to different types of diseases with different prognosis. Zembutsu et al. (2002) predicted the response to anti-cancer drugs in terms of efficiency and toxicity. Petty et al. (2006) discovered highly differentially expressed genes among cancer patients who responded to the chemotherapy treatment and those who did not. Along with the manifold increase in the quantity, and at the same time the quality of the measurements of the gene expression levels, there is an increasing need to study these measurements, which are followed by the statistical analysis.

The problem under study

There are many aspects in the statistical analysis of the microarray data that represent a great interest. One of the interesting problems, arising in the analysis of the microarray gene expressions is the choice of a model, the most suitable representation of the integrated behavior of gene expression levels. Discussions on the choice of the model of normal distribution, in other words, verification of the normality assumption of the gene expressions, has been discussed since the microarray technology was initiated and has been a big concern for many authors. Some authors proposed to use normality of the gene expression data (Stamey et al., 2001; Grant et al., 2002; Giles and Kipling, 2003; Guan and Zhao, 2005; Hardin and Wilson, 2009), while some others (Marko and Weil, 2012; Piras and Selvarajoo, 2015) provided more evidence against normality of the gene expressions rather than in favor. Thus, verifying normality assumption of the gene expression measurements is one of the most discussed issues in the literature. A detailed discussion of the research state on this issue, in connection with the question of the marginal normality of the logarithmic expressions, will be given in Chapter 1.

As for verifying the normality assumption, we propose some approaches for testing the normality of the gene expression data. Taking its origin from the problem of the large number of small samples, where each sample has its own parameters, so called nuisance parameters, our approach is based on the reconstruction of the type of distributions.

The measurements of the gene expression levels, produced by microarray technology, are aggregated expression signals and may not adequately reflect the events occurring within individual cells at the molecular level Chu et al. (2003). Therefore, even at the stage of designing a microarray experiment and measuring the expression levels, there are issues, such as the presence of the technical noise in the aggregate signals (Klebanov and Yakovlev, 2008; Hu et al., 2009), which should be taken into account, while proceeding with the analysis of the resulting measurements. Hence, there are a number of difficulties with the analysis of the measurements produced by microarray experiment. When proceeding with the statistical analysis of the measured levels of the gene expressions in general, and normality test in particular, we are faced with two main difficulties. The first one is the dependence between gene expressions and the existence of the diverse correlation structures between gene expressions Klebanov and Yakovlev (2007). The second difficulty is the small number of observations.

Because of these two difficulties we designed our study in the way that one can utilize the fact of having a large number of genes. Since gene expression data are highly correlated, therefore, we apply certain verified methods which make use of this fact. Introduced *a structure, yielding near-independent random variables*, subsequently called δ-sequences by Klebanov and Yakovlev (Klebanov and Yakovlev, 2007) is one of the most convenient techniques to deal with the dependence of the gene expression levels. One of the implications of δ-sequences, simply consists of transforming strongly dependent data into weakly dependent, or almost independent data. Hence, instead of testing normality of the original data, we test normality of the δ-sequences and apply the principle of the large number of small samples. A detailed discussion on this issue is represented in Chapters 1 and 4.

Structure of the work

This thesis consists of an introduction, four chapters and a conclusion.

Chapter 1 formulates the problem under study. In view of the importance of the methods of obtaining the gene expression data, this chapter presents some basics of the microarray technology in combination with the required background in molecular biology. The purpose of this is to show that the measured gene expressions data, due to the complexity of the microarray experiment, might contain certain "white noises", which interfere with the actual expression levels. A brief overview of some microarray techniques, the principle of microarray experiment, image processing and data acquisition are given. With a review of the related literature on normality issue, the problem of inter-gene dependence and the idea of using many genes are explained.

In Chapter 2 we discuss, mainly, the problem of the reconstruction of the type of distributions. In this chapter the problems of the large number of small samples and reconstruction of the type of distributions (additive, multiplicative and complete types) by the distribution of the maximal invariant statistics are explained. The problem of the characterization of the distribution is discussed. Several characterization theorems for the general case and two characterization theorems of the normal distribution are given. The last two characterization theorems were proven by Zinger (1956) and Sakata (1977a,b). These theorems provide a basis for replacing normality test with the spherical uniformity test. This chapter concludes by showing that characterization of the normal density given by Sakata does not hold for $k = 2$.

Chapter 3 deals with the test of spherical uniformity. The bases for the spherical uniformity test in Chapter 3 are characterization theorems from the previous chapter. For testing spherical uniformity in Chapter 3 we use statistics, developed by Bakshaev A. (Bakshaev, 2008, 2010). Since these types of statistics are based on $\mathfrak{N}$-distances, necessary information about $\mathfrak{N}$-distances is given. A testing procedure for the spherical uniformity test and its modification are described. Corresponding results of these tests are presented.

Chapter 4 represents the main part of our work. Results obtained in this chapter are new. In this chapter two characterization theorems are proven. The first theorem characterizes the normal distribution and the second is a characterization for the complete type of distribution. The first theorem can be considered as a specific case of the second one. Although there are results similar to these characterization theorems, these two theorems represent new results. These theorems are proven based on the analytic properties of the characteristic functions. These characterization theorems are, to some extent, similar to those in Chapter 2. Based on these theorems and by using Kolmogorov's test statistic, normality of the gene expressions is tested. By using two-sample Kolmogorov-Smirnov's test statistic, the relation of the two samples to the same type of distribution is tested. The results of the tests are presented at the end of the chapter.

With a brief overview of our study, Chapter 5 concludes the thesis.

Chapter 1

Formulation of the problem

1.1 Introduction

Before proceeding with a statistical analysis of any type of data, it is desirable to give some information on how these data are derived and what actual meaning they have. In this meaning statistical analysis in general, and normality test of the gene expression data in particular, deal with the same type of data, namely, with the measurements of the gene expression levels. The term of the gene expression data is referred to as the measured abundances of mRNA in a subset or in the complete set of genes in the genome of an organism. In other words, a gene, which is basically a protein coding, is considered to be expressed in a cell (or group of cells) when its transcribed messenger RNA (mRNA) is detected.

At the present day, there exist a wide range of different techniques to determine and to quantify the expression levels of genes, with substantial statistical components. The large-scale measurement of the gene expression levels is undergoing rapid development, which also has its statistical issues. In general, issues related to measuring the gene expressions or proteins are too broad, and this thesis has no intention to discuss these issues at length, rather to mention basic principles of quantifying RNA by extraction from a cell or tissue.

As was mentioned previously, in this thesis we study the validity of the assumption of whether the data of gene expression levels follow the normal law. With a brief introductory note on normality assumption of the gene expression data, this chapter formulates the problem under study. Due to the importance of the data acquisition methods for the analysis of the gene expressions, in this chapter we give a preview of some approaches, with focus on quantifying mRNA. Some basics of molecular biology with focus of microarray technology will be presented. Data sets that we dispose for the purposes of verifying normality assumption consist of a rather small number of slides (some tenth), each containing a large number (few thousands) of gene expressions. The structure of these data will be discussed later, in a relevant section.

1.2 Does the logarithm of the gene expressions follow the normal law?

A question of whether the gene expression levels or their logarithms follows the normal law came under discussion after the microarray technology was initiated and it turned out to be a big concern among many authors. As noted by Chen, Klebanov, Yakovlev (Chen et al., 2007), "The validity of the assumption on normality of gene expression measurements in microarray data has been a serious concern since the inception of this technology".

This question was a subject of investigation for many authors. In particular, a number of authors (Stamey et al., 2001; Grant et al., 2002; Giles and Kipling, 2003; Guan and Zhao, 2005; Lee et al., 2005b) studied this issue, giving preference to the use of nonparametric methods in microarray data analysis. Some other authors (Troyanskaya et al., 2002; Klebanov et al., 2006a; Lee et al., 2005a) also indicated certain advantages of nonparametric methods in such applications.

One of the first, if not the first, systematic studies, designed to test normality of the expression signals produced by all the genes in a microarray was undertaken by Giles and Kipling (Giles and Kipling, 2003). To verify normality assumption, the authors applied the Shapiro-Wilks test for normality to the expression levels of 12545 probe sets, produced by 59 human Affymetrix U95A GeneChips. The source of this data set is tissue from the human pancreas. The corresponding CEL files are accessible via Affymetrix website. According to the authors, for this data set the normality assumption is met and by that, the authors support the application of the parametric statistical tests which are based on the normality assumption of the gene expressions.

The study conducted by Chen, Klebanov and Yakovlev (Chen et al., 2007) shows that in order to verify the normality assumption, one has to have a larger data set, where both biological variability and technological noise are present. In order to systematically test for log-normality of the expression levels for all genes, they applied such test statistics as Kolmogorov, Cram'er von Mises, and Pearson χ^2 to a larger set of high-density oligonucleotide microarray data. For the non-normalized data, they did not reject the normality assumption of log-intensities. However, the global log-normality hypothesis was rejected for the data, normalized by the quantile normalization procedure. The results they obtained are consistent with the hypothesis that non-normalized expression levels of different genes are approximately log-normally distributed. From this it follows that the quantile normalization interferes not only with the technological noise but the true biological signal as well, and may radically change the marginal distributions of log-intensities.

By using 59 technical replicate in Affymetrix spike-in data set, Hardin and Wilson (Hardin and Wilson, 2009) investigated the assumption of normality of the oligonucleotide expression values. They concluded that transformed microarray data are not well-approximated by the normal distribution for any of the standard methods of calculating expression values. Taking into account the nature of microarray technologies, they presumed that this conclusion will be valid for other kinds of microarrays and they suggested that further study would be required to confirm the assumption of normality as it was already concluded in Chen et al. (2007).

To verify the validity of the normality assumption, Marko and Weil (Marko and Weil, 2012) designed a study based on cancer genomes. By using a variety of parametric and nonparametric methods, they concluded that cancer gene expression data are not normally distributed and they exhibit complex, heavy-tailed distributions, characterized by statistically-significant skewness and kurtosis.

To investigate transcriptome-wide variability of a single cells to different sizes of cell populations, Piras and Selvarajoo (Piras and Selvarajoo, 2015) examined RNA-Seq datasets of 6 mammalian cell types. They showed that for each cell type, increasing the number of cells reduces the variation in transcriptome-wide expressions and noise values and concluded that only the highly expressed portion of the genes in a single cells have Gaussian distribution.

As it follows from the above-mentioned studies, there is not a unique or predominant point of view regarding normality or non-normality of the gene expressions. Therefore, this question is still of great interest. In our study of verifying normality assumption of the gene expression data, we use an approach different from all that were mentioned above, based mainly on the characterization properties of the normal distribution. We carry out normality test by using two slightly different methods. In one of the methods, we replace the test of normality with the test of uniformity; instead of the normality test we conduct a test for the spherical uniformity. In the other method we just conduct a simple one-dimensional normality test of the gene expression data.

Because of the higher dimensions of the gene expression data, one may expect discussions on multivariate normality. But we emphasize that this thesis deals with one-dimensional normality test only.

1.3 A general overview of the logarithm of the gene expression data

This section gives a brief overview of the basic concepts involved in a microarray experiment and explains how gene expression levels are measured. Some computational methods that can be used to derive meaningful results from microarray experiments are described and some general information on logarithm of the gene expression data is given. For this purpose, the required biological background is presented; a definition of some biological entities is given and an explanation of the flow of genetic information within a biological system, which indicates the pathway of the gene expressions, is briefly described.

1.3.1 Biological background

To explain the microarray technology, a few biological terminologies from Alberts et al. (2015) are given below. The related sources for such biological background, necessary for an explanation of the gene expressions, are many, for example, Lodish et al. (2007); Bolsover et al. (2011).

- **Deoxyribonucleic acid (DNA)** is a molecule that consists of a long chain of nucleotides which in turn are composed of a nucleobase, the sugar deoxyribose and a phosphate. The nucleobases can be cytosine (C), thymine (T), adenine (A) and guanine(G). C and T are called pyrimidines and A and G are purines. DNA has a double-strand form, where each nucleotide binds its complementary nucleotide according to the pairing rule: A binds with T and G binds with C and vice versa. Discovered by Watson and Crick (1953), the DNA molecule has a well-known double-helix structure. The complete genetic information of an organism is stored in the nucleus of every cell in the form of double-stranded DNA that is curled up to build the chromosomes. DNA can be viewed as a long string from the letters A, C, G, T, denoting the four nucleobases. DNA and its building blocks are shown in Figure 1.1, which is taken from Alberts et al. (2015).

- **Ribonucleic acid (RNA)** is similar to DNA, but it has a single-stranded form and instead of ribose it contains sugar. Aside from this, in RNA the nucleobase thymine (T) is replaced by uracil (U). Depending on the functions, there are different types of RNA in the cell. A RNA which transports genetic information within the cell is called messenger RNA (mRNA).

- **Gene** is a segment of DNA that is transcribed as a single unit and carries hereditary information of a discrete hereditary characteristic. It usually corresponds to a single protein (or set of related proteins) or to a single RNA (or set of closely related RNAs).

- **Proteins** is the major macromolecular of the cell, its constituent, a linear polymer of amino acids linked together by peptide bonds in a specific sequence. Proteins have complex and versatile three-dimensional structure and performs many functions. It can be a constituent elements of the cell, enzymes or signaling molecules. Proteins can interact with other proteins and RNA's or DNA's. They make up more than half of the dry weight of the cell.

- **DNA polymerases** are enzymes that synthesize DNA (create DNA molecules) by joining nucleotides together. These enzymes are essential to DNA replication and usually work in pairs to create two identical DNA strands from a single original DNA molecule. During this process, DNA polymerase uses the existing DNA strands to create two new strands that match the existing ones. Every time a cell divides, DNA polymerase is required to help duplicate the cell's DNA, so that a copy of the original DNA molecule can be passed to each of the daughter cells. In this way, genetic information is transmitted to the next generation.

- **RNA polymerases** are called the enzymes that perform transcription. Like the DNA polymerase that catalyzes DNA replication, RNA polymerases catalyze the formation of the phosphodiester bonds that link the nucleotides together to form a linear chain. The RNA polymerase moves stepwise along the DNA, unwinding the DNA helix just ahead of the active site for polymerization to expose a new region of an template strand for complementary base-pairing. In this way, the growing RNA chain is extended by one nucleotide at a time in one direction.

- **Complementary DNA (cDNA)** is double-stranded DNA, synthesized from an mRNA template in a reaction catalyzed by the enzyme reverse transcriptase, an enzyme that catalyzes the formation of RNA from a DNA template during transcription, called RNA polymer. cDNA is often used to clone eukaryotic genes in prokaryotes. cDNA is usually used to express a certain protein in a cell that does not normally express such a protein. This process is referred to as heterologous expression. The expression of such a protein will be done by transferring the cDNA that codes for that protein to the recipient cell. cDNA can also be produced by retroviruses. Once the cDNA is created from such viruses, it is integrated into the genome of the host, where it goes on to create a provirus. When a protein is being synthesized, a gene's DNA is transcribed into an mRNA, which is then translated into a protein.

- **Translation (RNA translation)** is the process by which the sequence of nucleotides in an mRNA molecule directs the incorporation of amino-acids into protein. This process occurs on a ribosome.

- **Transcription (DNA transcription)** is the process of copying of one strand of DNA into a complementary RNA sequence by the enzyme RNA polymerase.

- **The cell** (from Latin cella, meaning "small room") is the basic structural, functional, and the minimal self-reproducing unit of all known living organisms. It consists of a self-replicating collection of catalysts. Reproduction of the cell (mitosis) is the transmission of genetic information to progeny cells. Every cell stores its genetic information in the same chemical form—as double-stranded DNA. The cell replicates its information by separating the paired DNA strands and using each as a template for polymerization to make a new DNA strand with a complementary sequence of nucleotides. A cell is separated from its environment by the plasma membrane. The membrane is the communication channel for a cell, it contains proteins that take up metabolites from the extracellular space or releases into that space. Cells can react to changes in their environment. Inside of the membrane the cell contains various organelles that serve different purposes. In principle, each cell consists of the same components. All cellular functions are governed by the information encoded in the genome, which is located in the nucleus of the cell.

  In order to serve special purposes, cells differentiate irreversibly. Due to gene expression modification, differentiation dramatically changes a cell's size, shape, metabolic activity, and responsiveness to signals. Cellular differentiation almost never involves a change in the DNA sequence itself (with a few exceptions). It involves switching off genes which are not needed in a particular tissue. Thus, despite having the same genome, cells in different tissues may have very different physical characteristics (Moore, 1972). In fact, cells differ in morphology (shape and appearance), metabolism (the complex of physical and chemical processes), gene expression and protein production. In terms of cellular regulation and dynamics, two cell types from one individual of the species can be as different as two unrelated bacteria.

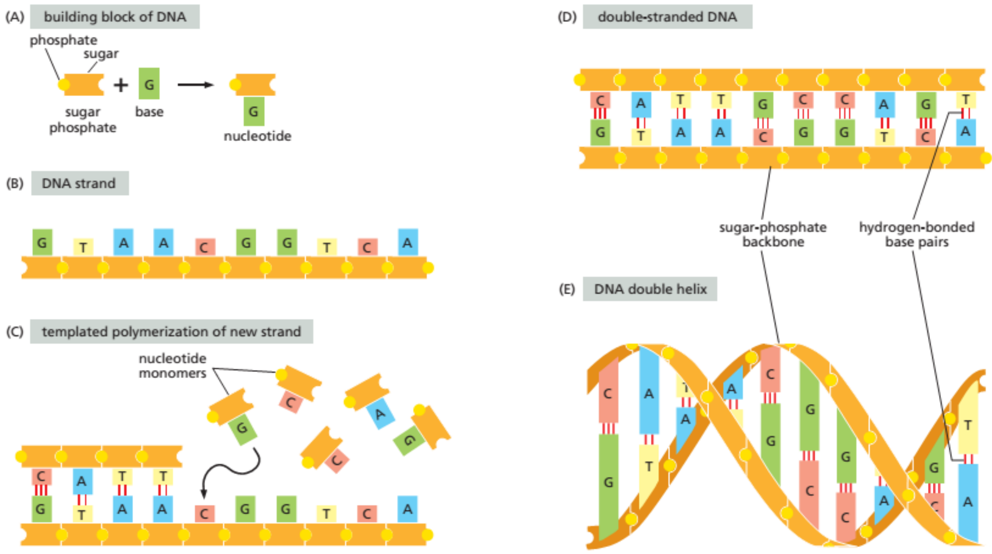


FIGURE 1.1: **DNA and its building blocks**. (A) DNA is made from simple subunits, called nucleotides, each consisting of a sugar-phosphate molecule with a nitrogen containing side group, or base, attached to it. The bases are of four types (adenine, guanine, cytosinecytosine, and thyminethymine), corresponding to four distinct nucleotides, labeled A, G, C, and T. (B) A single strand of DNA consists of nucleotides joined together by sugarphosphate linkages. Note that the individual sugar-phosphate units are asymmetric, giving the backbone of the strand a definite directionality, or polarity. This directionality guides the molecular processes by which the information in DNA is interpreted and copied in cells: the information is always "read" in a consistent order, just as written English text is read from left to right. (C) Through templated polymerization, the sequence of nucleotides in an existing DNA strand controls the sequence in which nucleotides are joined together in a new DNA strand; T in one strand pairs with A in the other, and G in one strand with C in the other. The new strand has a nucleotide sequence complementary to that of the old strand, and a backbone with opposite directionality: corresponding to the GTAA... of the original strand, it has ...TTAC. (D) A normal DNA molecule consists of two such complementary strands. The nucleotides within each strand are linked by strong (covalent) chemical bonds; the complementary nucleotides on opposite strands are held together more weakly, by hydrogen bonds. (E) The two strands twist around each other to form a double helix—a robust structure that can accommodate any sequence of nucleotides without altering its basic structure. This picture taken from Alberts et al. (2015)

.

According to the central principle of molecular biology (so called, the central dogma of molecular biology) DNA is transcribed into RNA which is, in turn, translated to produce proteins. DNA replication plays the central role in cell division.

The first step in producing protein is transcription of DNA by RNA polymerases, the enzymes that perform transcription. Then the resulted mRNA is translated into a chain amino-acids at the ribosomes, a cellular structure that is formed by ribosomal RNA and proteins. Translation takes place according to the genetic code, which is universal across all species and assigns each triplet of nucleotides one amino-acid. On a molecular level, the genetic code is realized with tRNA's, molecules of RNA that can recognize nucleotide triplets and bring the corresponding amino-acid to the ribosome, where it is attached to the growing protein chain.

Three well-known disciplines: genomics, transcriptomics and proteomics are related to the three central stages of molecular biology processes. Genomics is concerned with the structure, functions, evolution, and mapping of genomes, that is, with DNA; transcriptomics deals with the complete set of RNA transcripts that are produced by the genome, that is, mRNA; and finally, proteomics is connected with the structure, static description and dynamic behavior, and the analysis of the proteins occurring in living organisms Graves and Haystead (2002).

Proteins are the most active component in most cellular processes. Therefore, cellular processes are all about how proteins work and regulate. Translation of mRNA to protein is a very complicated process. If the relations were one-to-one, studying proteomics would be much easier. In that case, assuming that there is no regulation of protein degradation, the number of mRNA molecules of a particular gene would indicate the exact number of proteins after translation. But, unfortunately, there are some regulation mechanisms of the proteins concentration and their activities before and after the transcription. So, the combined consideration of all three -omics will result in a better understanding of cellular processes, regulation and functions.

1.3.2 Microarray Technology

Microarray technology allows one to study key biological questions on the genomic scale. It provides a systematic design for studying gene expressions, interpretation of the protein expression, discovery of molecular interactions, cellular functions and genetic studies as a whole. One of the most important applications of microarrays is transcriptional studies, measuring the entire repertoire of transcripts in a cell or in the whole organism, which provides a great deal of information on interactions between the DNA and the cellular phenotype. As noticed by Istepanian (Istepanian, 2003), to study the transcriptome comprising multiple transcripts expressed simultaneously in a cell or tissue, for example, about 300,000 RNA molecules in a human cell, is a challenge. In this meaning, microarrays are an extraordinary tool for the systematic analysis. Microarrays play an important role in the quantification of proteins expressions as well. In addition, microarrays are used in clinical studies, exploring the novelty of the technology in modern medical sciences. For example, microarrays were used successfully in cardiology to detect driver genes (Li et al., 2012).

A typical microarray can be defined as a small glass slide with a lot of spots or elements, where biochemical reactions take place. The microarray unfolds qualitative information on expression of gene or protein in a specific experiment. The word microarray comes from mikro (small) and arayer (arranged). A microarray has a rectangular shape and measures a few cm long; it is typically ordered, microscopic, planar and specifically coated with a suitable substrate (Schena, 2002)

Among others, there are two main technologies for manufacturing microarrays: spotted arrays and oligonucleotide arrays. Spotted arrays are produced by robotic spotting or by an inkjet printer and are referred to as cDNA microarrays or sometimes DNA microarrays. Oligonucleotide arrays are produced by one of the following approaches: photolithographic (Affymetrix array), inkjet technology (Agilent), electrochemical synthesis (CombiMatrix), solid state (NimbleGen), and silica beads in microwells (Illumina arrays; Illumina BeadArrays) (Smith et al., 2010; Arteaga-Salas et al., 2008; Draghici, 2012).

Microarrays, primeraly, were used to study the cellular processes by taking a snapshot of gene expressions. Microarrays discovery provided a common ground for collaborations among professionals in diverse disciplines such as biology, chemistry, toxicology, environmental studies, ecology, medical sciences, statistics, and computer science.

The common practice for using microarrays is to detect the presence of labeled nucleic acids in a hybridized with a probe nucleic acid biological sample, which is placed on the microarray surface. The detection of probes is promoted via bound labels that emit detectable by a laser scanner fluorescence (Schena, 2003). After image acquisition the analysis and interpretation of data follow, which provide a high-throughout analysis at the genomic scale. To produce large scale gene expression, the labeled nucleic acids are produced by reverse transcription expressed in a biological sample of mRNA. The biological sample might be from cells, tissues, or organisms under control or normal, and under treatment conditions (Hegde et al., 2000).

The labeling step depends on the experiment and the type of microarray technology to be used. In case of Affymetrix platform, a biotin-labeled complementary RNA target is constructed for hybridizing into the GeneChip. A common practice is the use of fluorescent labeling with two dyes Cy3 and Cy5 (detectable by green and red lasers).

In a typical experiment, two samples are hybridized to the arrays, each labeled with one dye, which allows the measurement of both fluorophores representing the nucleic acids expressed (or present) in each of two samples. This methodology allows the measurement of expression levels of many thousands of genes simultaneously. There exist many sources for the detailed description of the process of measuring gene expressions, microarray technologies, and microarray experiments, for example, Parmigiani et al. (2006); Amaratunga et al. (2014).

Below is a brief description of some of the microarrays techniques in molecular biology that conventionally have been used.

- **Southern blotting** (named after its discoverer Edwin Southern) is a procedure for identifying specific sequences of DNA in which fragments separated on a gel electrophoresis (which involves an electrical field) are transferred directly to a second medium on which assay by hybridization may be carried out. It comprises the electrophoresis of DNA molecules to generate smaller fragments. The fragments of a DNA which are separated electrophoretically are transferred and immobilized onto a solid support, usually a nylon or nitrocellulose membrane. In order to detect the presence of a particular DNA sequence on the membrane, a small fragment of DNA or an oligonucleotide corresponding to the target gene sequence is labeled with radioactive or non-radioactive fluorescent molecules. Subsequent to the hybridization of labeled probe with the membrane containing the immobilized DNA fragments, the detection of the bound fragments is carried out. This method allows one to identify the presence of a specific DNA molecule, which can be isolated and cloned for further analysis. However, the main limitation for the sequence detection using the Southern blot assay is that one can only use a single target gene at a time, although several samples are electrophoresed and immobilized on the solid support (Ali, 2014).

- **Northern blot** is commonly used to study transcriptional regulation of the gene expression by detecting specific RNA target molecules, corresponding to specific genes. The RNA is isolated from the cells, tissues, or organisms and is electrophoresed and transferred to a solid support, similar to DNA transfer in the Southern blotting procedure. The electrophoresis allows the separation of RNA molecules by size and they are detected by a hybridized labeled probe. The complementary target sequence is identified by detecting the bound labeled probe. Northern blotting is also limited to the detection of one gene at a time. However, multiple samples can be hybridized at the same time, and hence there is a possibility for the examination of cellular processes. Using northern blot allows one to routinely analyze regulation of gene expression during differentiation , morphogenesis, embryogenesis and development. Such analysis of gene expression regulation is possible in the cells under control, abnormal, or diseased conditions using the Northern blot procedure.

- **Western blot** or **immunoblot** detects specific proteins in a sample derived from cells, tissues, or organisms. The transfer of the protein samples is carried out similar to Southern and Northern blots. Electrophoresis of the protein sample separates the native proteins on an acrylamide gel. Based on the structure, weight, and charge of the native proteins which are present in the sample, the proteins may be electrophoresed using two or three-dimensional separation techniques. Then proteins are transferred to a membrane. Following the fixation of proteins on the solid support, the presence of specific polypeptide molecules is detected using specific antibodies, which are themselves detected using secondary antibodies labeled with fluorescent molecules. Due to the availability of numerous antibodies (monoclonal or polyclonal antibodies), detection of the expression of specific protein molecules in a given control or experimental sample becomes feasible. Limitation of the Western blot is the same as Southern and Northern blots: it allows a single type of protein to be detected in one experiment (Rueda, 2014).

- **Polymerase chain reaction (PCR)** is a cost-effective and time-saving technology, which gained popularity over Southern and Northern blots. It amplifies even a single copy of the molecule of a target nucleic acid in a biological sample. PCR is based on the exploitation of the ability of DNA polymerase to synthesize a new strand of DNA complementary to the offered template strand. Because DNA polymerase can add a nucleotide only onto a pre-existing three-prime-phosphate group, it needs a primer to add the first nucleotide. This requirement makes it possible to depict a specific template sequence region that should be amplified. At the end of the PCR reaction, the specific sequence will be accumulated in billions of copies (Rueda and Ali, 2014)

  To separate the strands from each other, at the beginning of the reaction a high temperature is applied to the original double-stranded DNA molecule. The most commonly used DNA polymerase is Taq DNA polymerase (an enzyme isolated from Thermis aquaticus bacteria), whereas Pfu DNA polymerase (from Pyrococcus furiosus) is used widely because of its higher fidelity in copying DNA. Despite some differences, these two enzymes have two capabilities suitable for PCR: 1) they can generate new strands of DNA using a DNA template and primers, and 2) they are heat resistant. The polymerase begins synthesizing new DNA from the end of the primer.

  Reverse Transcription PCR (RT-PCR) is a PCR preceded by the conversion of the sample RNA into cDNA with an enzyme. PCR and RT-PCR have the same limitations. The PCR reaction starts to generate copies of the target sequence exponentially. Only during the exponential phase of the PCR reaction it becomes possible to extrapolate back to determine the starting quantity of the target sequence, contained in the sample. Because of the inhibitors of the polymerase reaction, which are found in the sample, the chemical reagent limitation, accumulation of pyrophosphate molecules, and self-annealing of the accumulating product, the PCR reaction eventually ceases to amplify target sequence at an exponential rate and a "plateau effect" occurs, making the end point quantification of PCR products unreliable. This is the attribute of PCR that makes Real-Time Quantitative RT-PCR so necessary.

- **Real-Time Quantitative Reverse Transcription (qRT-PCR)** is a major development of PCR technology. It enables the reliable detection and measurement of products generated during each cycle of PCR process. This technique became possible after the introduction of an oligonucleotide probe which was designed to hybridize within the target sequence. The split of the probe during PCR because of the five-prime-phosphate group (5') nuclease activity of Taq polymerase can be used to detect amplification of the target-specific product. A technique to monitor degradation of the probe is the implementation of double-stranded DNA-binding dyes. Probe labeling with fluorescent dyes is used as well.

  One of the earliest methods introduced for qRT-PCR monitoring was TaqMan assay (named after Taq DNA polymerase). This method has been widely used for the quantification of mRNAs and for detecting its variation. The method exploits the five-prime-phosphate group (5') endonuclease activity of TaqDNA polymerase to split an oligonucleotide probe during PCR, thereby generating a detectable signal. The probes are fluorescently labeled at their five-prime-phosphate group (5') end and are non-extendable at their three-prime-phosphate group (3') end by chemical modification. Specificity is conferred at three levels: via two PCR primers and the probe. Real Time Quantitative RT-PCR is used for the relative and absolute quantification of gene expressions and validation of DNA microarray results.

All above-mentioned microarray techniques are the low-throughout methods. What unites them is that all of them are based on hybridization (Speed, 2004). High-throughout methods are Serial Analysis of Gene Expression (SAGE) and array based approaches. SAGE is a method for the comprehensive analysis of gene expression patterns and it not based on hybridization. Besides, one of the advantages of SAGE is that it is not necessary to know the sequences of the mRNA transcripts in advance. There are three main principles in SAGE:

1. A short sequence tag (10-14bp) contains sufficient information to uniquely identify an mRNA transcript, provided that that the tag is obtained from a unique position within each transcript.
2. Sequence tags can be linked together to form long serial molecules that can be cloned and sequenced.
3. Quantification of the number of times a particular tag is observed provides the expression level of the corresponding transcript.

A typical SAGE experiment involves two sources of mRNA. For each source a set (library) of tags would be derived using the SAGE protocol. In these two libraries there might be many distinct tags observed, and for each unique tag, the frequency of the tag appearances in each library could be calculated. The data for this comparative experiment are then two lists of counts, one for each unique tag observed.

Array based approaches are principal part of the high-throughput methods for quantifying gene expression. There are three basic microarray technologies: nylon membrane arrays, spotted arrays, and high-density oligonucleotide arrays.

Nylon membrane Filters is the oldest array technology, but still is widely used around the world. A typical filter microarray has 5000 complementary DNA (cDNA) clones 600-2400 bases in length, spotted in a grid on the membrane. Radio-labeled target cDNA derived from the mRNA of interest is hybridized to the array, and the filter is then exposed to X-ray film and the film imaged. The digital image is the raw data from the experiment. Traditional high-density filter-based microarray is the oligonucleotide filter array, which can have 50,000 spots (Meier-Ewert et al., 1998).

Spotted cDNA Microarrays was introduced by Schena (Schena, 1995). It is a typical spotted array, which consists of 40,000 cDNA probes with the length of 600-2400 bp placed in a regular pattern on a glass microscope slide.

High-density oligonucleotide arrays is quite different technology that can place up to 500,000 short oligonucleotide probe pairs on a small glass chip, with 11-20 of them representing part or all of a single gene (Lockhart et al., 2000; Fodor et al., 1991). Each probe pair consists of a perfect match (PM) probe, and a mismatch (MM) probe. MM is the same as PP aside from a difference in a single nucleotide change in the middle (13th) position. A tagged target cRNA sample hybridizes with the complementary oligonucleotides on the chip, and detection is via laser excitation followed by the collection of fluorescence emission as with spotted arrays.

The end product of the experimental stage is an image of the microarray, where each spot that corresponds to a gene, has an associated fluorescence value representing the relative expression level of that gene. The main starting point in microarray data analysis is to process this image, which involves several steps (Babu, 2004a).

1. Spot identification and distinguishing them from spurious signals. Scanned microarray is followed by hybridization and usually a TIFF image file is generated. Once image generation is completed, the image is analysed to identify spots. To make spot identification straightforward, in microarray images the spots are arranged in an orderly manner into sub-arrays. To identify regions corresponding to spots, image processing requires layout specification of each sub-array.

2. Determination of the spot area and the local region to estimate background hybridization. After identifying corresponding regions to sub-arrays, an area within the sub-array is selected in order to get a measure of the spot signal and background intensity is estimated. There are two methods to define the spot signal: using a fixed size area, centered on the centre of the spot mass, which is computationally less expensive but with some disadvantages; precise definition of the boundary for a spot that includes only pixels within the boundary, which gives a better estimation of the spot but is time-consuming and computationally expensive.

3. Summary statistics and assigning spot intensity after correction for background intensity. Once the spot and background areas are defined, a variety of summary statistics for each spot (red and green) are reported. Each pixel is taken into account and summary statistics (mean, median, total) for the intensity in all the pixels of the defined area for both the spot and background are reported. Some approaches use the spot median value, subtracted from the background median value as the metric to represent spot intensity. The advantage of this approach is its relative insensitivity to a few pixels with anomalous fluorescent values and its disadvantage is its sensitivity to misidentification of spot and background areas. Another approach uses total intensity values, which has an advantage of being insensitive to misidentifcation of spots but has a disadvantage of being inclined to be skewed by a few pixels with extreme intensity values.

One of the important considerations in image processing is the choice of the number of pixels to be included for measurement in the spot image (Babu, 2004b). For many scanners, the default pixel size is $10\mu m$. This means that an average spot of diameter of 200 μm will have approximately 314 pixels. However, for a smaller spot diameter, it is better to use a smaller pixel size to ensure enough pixels are sampled. Most scanners now allow much smaller pixel sizes but the size of the image file increases.

1.3.2.1 Logarithmic transformation

Often the data of spot intensity is initially transformed for analysis by a logarithmic transformation, $X \to \log(X)$ (Colantuoni et al., 2003). It is preferable to work with log-intensities rather than absolute intensities for a number of reasons: the variation of log-intensities tends to be less dependent on the magnitude of the values, taking logarithm reduces the skewness of highly skewed distributions, and improves variance estimation. Furthermore, log-intensities elevates visual inspection of the data. Often the raw data is heavily clumped together at low intensities followed by a very long tail, the details of such configurations are impossible to discern. Log transformation spreads out the data more evenly, making it easier to examine visually. Often logarithms of base 2 are used. Other simple power transformations (of the form $X \to X^b$ for some $b > 0$) have been found to be useful for certain datasets. For example, Amaratunga and Cabrera (2001b,a) use a square root transformation: $X \to \sqrt{X}$, (Tusher et al., 2001) uses a cube root transformation: $X \to X^{1/3}$. But the log transformation is the most widely used.

Thus, through multiple-stage procedures of experiment, image acquisition, a few intermediate steps for normalization, background correction, etc., and finally the log transformation, we obtain a matrix of the log- intensities of the gene expression levels of the form

$$\mathbb{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \cdots & \cdots & \cdots & \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix},$$

where element X_{ij} denotes the j-th expression level of the i-th gene.

1.4 Dependence between gene expressions

One of the issues that should be addressed in the analysis of the gene expression data is the dependence structure of the gene expressions. A comprehensive study of the diverse dependence structures of the gene expressions was designed by Klebanov and Yakovlev (2007). By analyzing various data sets, they observed that the average of the correlation coefficients, calculated over all gene pairs, varies from 0.84 to 0.97.

A type of the correlation structure of the microarray data that they studied can be considered as a long-ranged property of the correlations, meaning that a particular gene may have very high correlation coefficients with a vast majority of other genes. In a typical example by a particular gene IPF1 in PCMIT that encodes a transcription factor involved in regulation of transcription and morphogenesis, they showed that the mean value of the correlation coefficient is 0.78, while the corresponding standard deviation is equal to 0.16. Such long-range strong correlations prevail in a huge proportion of randomly selected genes. They state that to make the analysis of correlations tractable, it is important to identify patterns of stable correlation and patterns that are either universal or specific to the phenotype under study.

In another paper by Klebanov, Jordan and Yakovlev Klebanov et al. (2006b), a special type of stochastic dependence between expression levels in pairs of genes is described. A modulation-like unidirectional dependence between expression signals was studied in three large sets of microarray data on childhood leukemia. Later on, this type of dependence was confirmed by a similar analysis of some other data sets. This type of dependence in (Klebanov et al., 2006b) is termed as the type A dependence. A distinctive feature of this type of dependence is that the expression of a “gene-modulator” is stochastically proportional to that of a “gene-driver”. A formal definition of type A dependence can be described as follows. Let g_x and g_y be two given pair of genes, and let random variables X and Y represent their respective expression levels. A pair of genes (g_x, g_y) is said to be type A, if X and Y satisfy the condition

$$Y = XZ, \tag{1.1}$$

where Z is a positive random variable, stochastically independent of X. In this case, g_x is called a driver and g_y a modulator. Introducing this terminology is reasoned by non-symmetrical roles of $g_x$and g_y in a type A pair. If (1.1) is true, the random variables Y and $1/Z$ are no longer independent; this is precisely what makes the type A stochastic dependence so special. All other stochastic dependencies in gene pairs are classified as type B dependence. Log-transforming of the expression (1.1) gives

$$y = x + z, \tag{1.2}$$

where $x = \log X$, $y = \log Y$ and $z = \log Z$. A necessary condition for the type A dependence is

$$Var(x) = Cov(x, y). \tag{1.3}$$

This is a sufficient condition under *joint normality* of (x, y). Since $Var(x) < Var(y)$ in each type A pair genes, then from (1.2) it follows that the type A dependence induces an ordering of the random variables x and y in terms of their variances.

The presence of a multiplicative technical noise is the main concern in the studies of the correlation structure of microarray data that can induce spurious correlations.

Another interesting phenomenon has been discovered by Klebanov, Jordan and Yakovlev (Klebanov et al., 2006b) in triples of the gene expressions. The between-gene correlations are overwhelmingly positive. An essential part of strong positive correlations is contributed by type A dependence in numerous gene pairs. If we consider a triple of genes formed by two type A pairs, we observe that the vector of expressions with arranged variances in increasing order is of the form $(x, x + z_1, x + z_1 + z_2)$, where z_1 is independent of x and z_2 is independent of $x + z_1$. Therefore, it follows the following relationship

$$Cov(x, z_2) = -Cov(z_1, z_2). \tag{1.4}$$

This shows that the covariance between the increments z_1 and z_2 is expected to be negative whenever $Cov(x, z_2) > 0$. How frequently the latter condition is met can be assessed only with real data.

In many studies of microarray data, authors use some independence assumptions across genes. Such assumptions can result in convenient mathematical forms, simplify the question under the study, for example, estimating the proportion of equivalently expressed genes (Benjamini and Hochberg, 1995; Storey, 2002; Nettleton et al., 2006). However, as mentioned above, there exist strong correlations between gene expression. This is from one side. From another side, genes are related to each other intricately through regulatory networks (Altman and Raychaudhuri, 2001; Wyrick and Young, 2003), resulting in correlations between measured expressions.

1.5 The idea of using many genes

In the previous section, the issue of dependence between gene expressions, the existence of diverse correlations structures in the gene expression data was explained. Another issue that adds more burden while analyzing gene expression data is the number of observations. Samples of the gene expression data are usually small, ranging from 5 to 20, only in some cases around 100. Since the number of genes themselves is large enough, we can hope to take advantage of it.

Before discussing the sample size, we continue with the issue of dependence of the gene expressions. If there was not dependence between gene expressions, then dealing with them would be easier and the analysis of the mciroarray data as a whole could probably be conducted in a more simplified form. For example, knowing that the k-th gene does not depend on any other gene or there is not any type of correlations between the k-th and all other measured expression levels, we could consider n observations for the k-th gene as a one-dimensional sample of size n. But as demonstrated above, strong correlation has been evidenced by many research papers. Thus, by now the question should not be about dependence but rather how to deal with inter-gene dependence.

We refer again to the paper by Klebanov and Yakovlev (2007), where they introduced “a structure yielding near-independent random variables”. The idea of this structure consists of pooling gene expression levels (or associated test statistics) across genes in such way that it will result in better procedures in the analysis of the gene expressions. Originally, this idea was developed in Qiu et al. (2005) and Klebanov and Yakovlev (2006) for selecting differentially expressed genes and the purpose was in some way to compensate the small sample size due to a large number of genes. The authors of these papers relied on the fact that the number of genes m is very large, so that the effective sample size can be increased and the asymptotics in the sample size n can be replaced by the asymptotics in the number of genes m. If gene expression data or some statistics derived from them were independent, then this idea would work well even with small samples. But as stated in Klebanov and Yakovlev (2007), the actual correlation structure of the gene expressions stands in the way of pooling across genes.

If genes are ordered by increasing variances (each gene is assigned a number from $i = 1$ through $i = m$, where $i = 1$ corresponds to a gene with the minimal variance, while $i = m$ corresponds to a gene with the maximal variance), then the pairwise correlations in such ordered sequences are very high. For example, for the genes with even numbers in all estimated pairwise correlation coefficients between their expression levels, strong positive correlations will prevail. The mean value of these correlation coefficients is 0.942 with the corresponding standard deviation being equal to 0.036.

The situation changes radically if one constructs the sequence of increments $\delta_i = x_{2i} - x_{2i-1}$ between log-expression levels of the $2i$-th and the $(2i - 1)$-th gene $(i = 1, \dots, m)$ in the above-defined ordering of genes. This specific sequence δ_i, henceforth termed the δ-sequence, generates near-independent random variables.

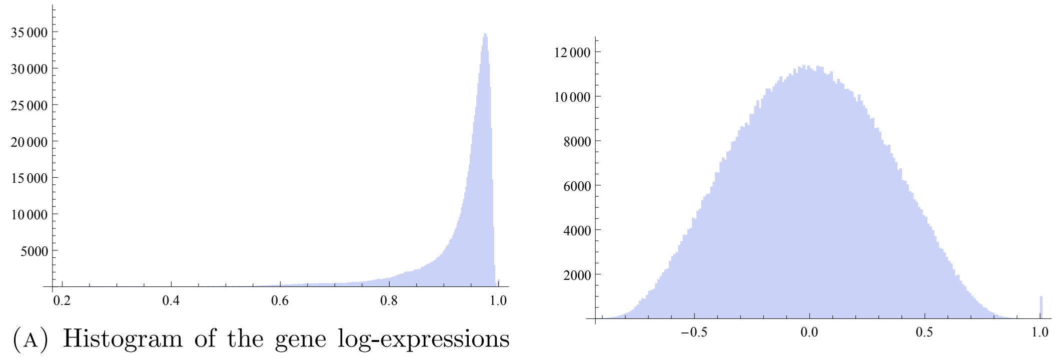
In our work we use the δ-sequence. As in the paper by Klebanov and Yakovlev (2007), we first arrange gene expressions by increasing variance and then take the differences of the two neighborhood genes. Due to weak dependence, δ-sequences are considered as independent random variables. Then, for the data of each separate gene expression, we will have a (one-dimensional) sample of the same size as that of the original data set of the gene expression levels. But due to pairing, the number of samples (genes) will be $[m/2]$, where m is the number of the expressed genes, $[a]$ denotes the integer part of a.

Histograms of correlation coefficients of the gene expressions and corresponding δ-sequences are shown in Figure 1.2. The Figure 1.2 compares the sample of correlation coefficients between log-expression levels with those observed in pairs of the δ sequences. While in the former case the histogram of correlation coefficients is shifted to the right (Figure 1.2a), it becomes symmetric around zero with the increments, δ sequences (Figure 1.2b). Similar properties of the δ-sequences versus gene log-expressions have been confirmed for both data sets HYPERDIP and TELL.

1.5.0.2 Human genome

Since this thesis deals with data of gene expressions obtained from human tissues (patients with leukemia), we will give a short overview of the human genome.

Eukaryote is an organism composed of one or more cells that have a distinct nucleus. It is a member of one of the three main divisions of the living world, the other two being bacteria and archaea (Alberts et al., 2015). The human organism, like many other beings such as animals, plants, etc., belongs to eukaryotes, whose cells contain a nucleus and other organelles enclosed within membranes. Most eukaryotic organisms have billions of



(A) Histogram of the gene log-expressions data

(B) Histogram of the δ sequences

Figure 1.2: Histogram of correlation coefficients formed by all pairs of 1000 randomly chosen gene expressions from HYPERDIP data set; (a) is the histogram of the correlation coefficients, calculated for all pairs of 2500 gene log-expressions; (b) is the histogram of the corresponding δsequences. The mean value of the correlation coefficients for the log-expression is 0.929766 and for δ sequences this number is equals to 0.00116495. The corresponding minimal values for log expressions and δ-sequences are 0.182785 and -0.92003, respectively. Similar picture we observe for the TELL data set.

individual cells. Almost all of these cells contain the entire genome for that organism. This genome carries complete hereditary information in the form of DNA.

The human genome consists of 23 pairs of chromosomes. Each chromosome is made up of chains of DNA. In humans, there are about 27,000 genes. Some statistics of the human genome are shown in Table 1.1 Alberts et al. (2015).

1.5.0.3 Data used in the thesis

In this thesis, two data sets of the gene expressions, called HYPERDIP and TEL data sets for childhood leukemia are used. These data sets are accessible via the website of St.Jude Chilren's research hospital, http://www.stjude.org. These data sets were processed by Affymetrix microarrays; both consist of expression levels of 7084 genes and neither of them is normalized. HYPERDIP data set has 88 and TEL data 79 observations (slides). Detailed information on processing can be found in supplementary materials of Yeoh et al. (2002).

Table 1.1: Some statistics of the human genome

Human Genome	Statistics
DNA length	3.2 × 109 nucleotide pairs [a]
Number of genes coding for proteins	Approximately 21,000
Largest gene coding for protein	2.4 × 106 nucleotide pairs
Mean size for protein-coding genes	27,000 nucleotide pairs
Smallest number of exons per gene	1
Largest number of exons per gene	178
Mean number of exons per gene	10.4
Largest exon size	17,106 nucleotide pairs
Mean exon size	145 nucleotide pairs
Number of noncoding RNA genes	Approximately 9000 [b]
The number of pseudogenes[c]	More than 20,000
Percentage of DNA sequence in exons (protein-coding sequences)	1.5 %
Percentage of DNA in other highly conserved sequences[d]	3.5 %
Percentage of DNA in high-copy-number repetitive elements	Approximately 50 %

[a] The sequence of 2.85 billion nucleotides is known precisely (error rate of only about 1 in 100,000 nucleotides). The remaining DNA primarily consists of short sequences that are tandemly repeated many times over, with repeated numbers differing from one individual to the next. These highly repetitive blocks are hard to sequence accurately.

[b] This number is only a very rough estimate.

[c] A pseudogene is a DNA sequence closely resembling that of a functional gene, but containing numerous mutations that prevent its proper expression or function. Most pseudogenes arise from duplication of a functional gene, followed by the accumulation of damaging mutations in one copy.

[d] These conserved functional regions include DNA encoding 5' and 3' UTRs (untranslated regions of mRNA), DNA specifying structural and functional RNAs, and DNA with conserved protein-binding sites.

Chapter 2

Reconstruction type of distributions

2.1 Introduction

Having taken its origin from the problem of the large number of small samples with nuisance parameters, this chapter deals with the problem of reconstruction of the original (or parent) distribution by the distribution of its maximal invariant statistic. The materials for this chapter are based on results obtained by Linnik (1956), Zinger (1956), Kovalenko (1958), Prokhorov (1965), Zinger and Linnik (1964), Zinger and Kagan (1976), Kagan and Zinger (1977), Kakosyan et al. (1984).

2.2 Reconstruction the type of distributions

Fairly often in statistical practice, it becomes necessary to test a hypothesis concerning whether or not a distribution belongs to a given family of distributions, by a number of small samples. When testing this hypothesis, the parameters of the distribution, in general, may change from sample to sample and therefore, for testing this type of hypothesis one has to use statistics whose distributions, in a certain sense, eliminate these parameters.

Thus, there arises an analytical problem of reconstruction of the original distribution by the distribution of some statistic. In statistical literature there exists a considerable number of papers and monographs devoted to the study of this problem. Among others, one can mention Kagan et al. (1973), Galambos and Kotz (1978), Kakosyan et al. (1984), in which enough detailed information and references are given. By using the concept of intensively monotone operators in Kakosyan et al. (1984) a wide range of characterization theorems are given, in particular, for the normal distribution.

In general, the problem of reconstruction of the type of distribution can be formulated as follows. Let $x_1^{(j)}, \ldots, x_{n_j}^{(j)},\ j = 1, \ldots, m$ be m samples of sizes n_j. The question is how to verify that all samples belong to same type of distribution.

For the case when $x_i^{(j)},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m$ belong to the normal distribution with the mean a_j and the variance σ_j^2, the values of a_j and σ_j^2 are considered as the nuisance parameters. To eliminate them one considers the studentized differences

$$y_i^{(j)} = \frac{x_i^{(j)} - \bar{x}^{(j)}}{S_j}, \quad S_j^2 = \sum_{i=1}^{n_j} (x_i^{(j)} - \bar{x}^{(j)})^2, \tag{2.1}$$

which under the null hypothesis that all samples belong to the normal type, are distributed uniformly on the unique sphere. A detailed study of the problem with nuisance parameters is given in Linnik (1968).

The procedure of eliminating the nuisance parameters by applying the studentized differences is referred to as the problem of large number of small samples. We begin with the reconstruction of the additive type of distributions (see Kovalenko (1958); Prokhorov (1965); Kagan et al. (1973)).

2.2.1 Reconstruction of the additive type

Definition 2.1. We say that a distribution belongs to the additive type if its density is given by

$$p(x, \theta) = p(x - \theta), \tag{2.2}$$

where $\theta \in \mathrm{R}^1$ is a translation parameter.

Let $X_1, \ldots, X_n$ be a sample from a distribution given by (2.2). Put

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

and consider a random vector

$$\mathrm{Y} = (X_1 - \bar{X}, \ldots, X_{n-1} - \bar{X}). \tag{2.3}$$

Evidently, the distribution of the random vector Y does not depend on θ, that is, the statistic Y eliminates it. A natural question which arises in this case is: to which extent does the distribution of the statistic Y determine the distribution of the original sample?

Under some additional assumptions, by the distribution of the statistic Y one can determine the distribution of the original sample within a translation parameter (Kovalenko, 1958). More precisely, for $n \geq 3$ distribution of Y determines the distribution of X_1, or in other words, the characteristic function

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} p(x) dx$$

up to a factor $e^{i\gamma t}$ in the case when $\varphi(t) \neq 0$.

A counter-example was given, which shows that if $\varphi(t)$ has zeros, then the reconstruction of the additive type is impossible. A more general result on reconstruction of the complete type of distribution was obtained by Prokhorov Yu.V. It was proven that for $n \geq 6$, under some assumptions, the distribution of X may be reconstructed from the distribution of Y up to location and scale parameters (Prokhorov, 1965).

In principle, as the statistics Y one can take any single-valued invertible function, for example, $n-1$- dimensional vector

$$\mathrm{Y}' = (X_1 - \bar{X}, \ldots, X_n - \bar{X}),$$

which is the maximal invariant statistic.

Statistic Y given by (2.3) eliminates the translation parameter, that is, the distribution of the random vector Y determines the additive type. Hence, to verify that two distributions with densities $p(x)$ and $q(x)$ differ only by a translation parameter, one can verify the coincidences of distributions of the statistics Y, induced by $p(x)$ and $q(x)$ under some conditions. But, statistic Y is a vector of dimension $n-1$ and verifying coincidence of multivariate distributions is not an easy task. Therefore, it is desirable to pass from statistic Y to some one-dimensional statistic, more preferably to some linear one, since Y is a linear statistic.

2.2.2 Reconstruction of the multiplicative type

Definition 2.2. We say a distribution belongs to scale type if its density is given by

$$p(x,\theta) = \frac{1}{\theta} p\left(\frac{x}{\theta}\right), \tag{2.4}$$

where $\theta > 0$.

This type can simply be reduced to the additive type by taking logarithm of the random variables and changing parameters, where $\tilde{\theta} = \log\theta$ will play the role of the translation parameter in the additive type.

If $X_1, \ldots, X_n$ is a sample of size n from a distribution function, defined by (2.4), then distribution of the two-dimensional random variables

$$Y_i = (\log|X_1|\,,\ \mathrm{sign}\, X_i)$$

belongs to the additive type with the density

$$q(y, \tilde{\theta}) = q(y_1 - \tilde{\theta}, y_2),$$

where $\tilde{\theta} = \log\theta$.

2.2.3 Reconstruction of the complete type

Definition 2.3. We say a distribution belongs to complete type if its density is given by

$$p(x,\theta) = \frac{1}{\sigma} p\left(\frac{x-a}{\sigma}\right), \tag{2.5}$$

where $\theta = (a, \sigma),\ a \in \mathrm{R}^1$ and $\sigma > 0$.

For this type the maximal invariant statistic Y is a $(n-2)$-dimensional vector

$$\mathrm{Y} = (y_1, \ldots, y_n),$$

where

$$y_k = \frac{x_k - \bar{x}}{s}, \quad s^2 = \sum_{k=1}^{n} (x_k - \bar{x})^2, \quad s > 0.$$

It is clear that $\sum_{k=1}^{n} y_k = 0$ and $\sum_{k=1}^{n} y_k^2 = 1$, that is, $(y_1, \ldots, y_n)$ belongs to $n-2$-dimensional sphere.

For $n \geq 6$ Zinger and Kagan (1976), under less restrictive assumptions than in Prokhorov (1965), gave sufficient conditions for reconstructing the complete type by the distribution of the maximal invariant statistic.

2.3 Characterization of the normal density

In this section we give some characterization theorems of the normal density which will be used in subsequent chapters.

2.3.1 Characterization theorem by A. A. Zinger

Let $x_1, x_2, \ldots, x_n$ be independent and identical distributed (i.i.d) random variables. Consider statistics

$$z_1 = \frac{x_1 - \bar{x}}{S},\ z_2 = \frac{x_2 - \bar{x}}{S}, \ldots, z_n = \frac{x_n - \bar{x}}{S}, \tag{2.6}$$

where

$$n\bar{x} = x_1 + x_2 + \ldots + x_n, \quad S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Then the random vector $\mathbb{Z} = (z_1, z_2, \ldots, z_n)$ is distributed on $n-2$-dimensional sphere

$$\Phi^{n-2} = \begin{pmatrix} z_1 + z_2 + \ldots + z_n = 0 \\ z_1^2 + z_2^2 + \ldots + z_n^2 = 1 \end{pmatrix}.$$

Clearly, the distribution of random variables $x_1, x_2, \ldots, x_n$ defines the distribution of the random vector $\mathbb{Z}$. Moreover, if random variables $x_1, x_2, \ldots, x_n$ have normal distribution, then the random vector $\mathbb{Z}$ is distributed uniformly on the sphere Φ^{n-2}. The inverse problem was solved by Zinger A. A. for $n \geq 6$.

The inverse problem, in general case, was set by A. N. Kolmogorov. Zinger solved it for the case of normal density. Namely, the following theorem was proved by him.

Theorem 2.3.1 (Zinger (1956))**.** *If the random vector $\mathbb{Z}$ has a uniform distribution on the sphere Φ^{n-2} and $n \geq 6$, then random variables $x_1, x_2, \dots, x_n$ have normal distribution.*

Formally, Zinger's theorem may be obtained from Prokhorov's theorem, but it was obtained 3 years before, and by another method.

The characterization theorem by Zinger was significantly strengthened by (Zinger and Linnik, 1964). It was shown that for the characterization of the normal density, uniformity of the distribution of the statistics $z_1, \dots, z_n$ on the entire sphere Φ^{n-2} was not necessary; it is sufficient if the density is the same for a finite set of points on the sphere. Precisely, the following theorem was proved.

Theorem 2.3.2. *Let $X_1, \dots, X_n$ be a sample from a one-dimensional distribution with a continuous density $f(x)$. If on the sphere Φ^{n-2} could be found at least one triplet of points $\boldsymbol{y}, \boldsymbol{y}', \boldsymbol{y}''$ of the form*

$$\begin{aligned}
\boldsymbol{y} &= (y_1, y_2, y_3, y_4, y_5, y_6, y_7, \dots, y_n), \\
\boldsymbol{y}' &= (y_1, y_2, y_3, y_1, y_2, y_3, y_7, \dots, y_n), \\
\boldsymbol{y}'' &= (y_4, y_5, y_6, y_4, y_5, y_6, y_7, \dots, y_n)
\end{aligned}$$

such that

$$\begin{aligned}
y_1 + y_2 + y_3 &= y_4 + y_5 + y_6, \\
y_1^2 + y_2^2 + y_3^2 &= y_4^2 + y_5^2 + y_6^2,
\end{aligned}$$

but

$$(y_1, y_2, y_3) \neq (y_4, y_5, y_6),$$

for which the density of the vector $(y_1, \dots, y_n)$ on the sphere Φ^{n-2} is the same, then the distribution of the original sample is normal.

2.3.2 Characterization theorem by T. Sakata

Let $(X_{i1}, X_{i2}, \dots, X_{i2k})$, $i = 1, \dots, n$ be a sequence of samples with size $2k$, being independently drawn from a population Π_i with a continuous density function

$$\frac{1}{\sigma_i} p\left(\frac{x - \mu_i}{\sigma_i}\right),$$

where $p(.)$ is a symmetric density and $-\infty < \mu_i < +\infty, \quad \sigma_i > 0, \quad (i = 1, \dots, n)$.

Sakata studied the problem of reconstructing the common density $p(.)$ where the unknown parameters (μ_i, σ_i) may change from one population to another, and the sample size, $2k$, is so small that the estimate of the unknown parameters are not available with enough accuracy. Therefore, he proposes a series of transformations to eliminate the unknown parameters (μ_i, σ_i), which could be expressed as follows

$$Z_j = \frac{|Y_j|}{\sqrt{\sum_{j=1}^{k} Y_j^2}}, \tag{2.7}$$

where $Y_j = X_{2j-1} - X_{2j},\ j = 1, \dots, k$.

Theorem 2.3.3 (Sakata (1977b)). *Let statistics $Z_j,\ j = 1, \dots, k$ are defined by (2.7).*

1. *If $h(\mathbb{Z})$ is the density function of the statistic $\mathbb{Z} = (Z_1, Z_2, \dots, Z_k)$, which takes value on the set*

   $$S_+^{k-1} = \left\{ \sum_{j=1}^{k} Z_j^2 = 1,\ Z_j \geq 0,\ j = 1, \dots, k \right\},$$

   then

   $$h(\mathbb{Z}) = c \int_0^{\infty} s^{k/2-1} f(\sqrt{s Z_1}) \dots f(\sqrt{s Z_k}) ds, \tag{2.8}$$

   where $f(.)$ is a convolution of $p(.)$ and c is a constant.

2. *If $p(.)$ is the standard normal density, then the random variable $\mathbb{Z} = (Z_1, Z_2, \dots, Z_k)$ is uniformly distributed on S_+^{k-1}.*

3. *If the density of $p(.)$ has a differentiable bounded convolution and for $k \geq 3$ statistic $\mathbb{Z} = (Z_1, Z_2, \dots, Z_k)$ is uniformly distributed on S_+^{k-1}, then $p(.)$ is the density of standard normal distribution.*

From (2.7) by using spherical coordinates

$$\begin{aligned}
&Y_1 = \rho \cos \varphi_1 \\
&Y_2 = \rho \sin \varphi_1 \cos \varphi_2 \\
&Y_3 = \rho \sin \varphi_1 \sin \varphi_2 \cos \varphi_3 \\
&\dots\dots\dots\dots \\
&Y_{k-1} = \rho \sin \varphi_1 \cdots \sin \varphi_{k-2} \cos \varphi_{k-1} \\
&Y_k = \rho \sin \varphi_1 \cdots \sin \varphi_{k-2} \sin \varphi_{k-1},
\end{aligned}$$

where

$$\rho \geq 0, \quad 0 \leq \varphi_j \leq \pi,\ j = 1, \dots, k-2,\ 0 \leq \varphi_{k-1} \leq 2\pi,$$

we obtain

$$\rho Z_j = Y_j,\ j = 1, \dots, k. \tag{2.9}$$

2.3.2.1 Case $k = 2$

Following a counter-example by Laha R. (Laha, 1958), where the quotient of two non-normal random variables follows Cauchy distribution, one can show that the result obtained by Sakata for $k = 2$ does not hold.

Let $k = 2$ and $f(y_1)f(y_2)$ be the joint density of 2-dimensional random vector (Y_1, Y_2). Then change to spherical coordinates gives

$$f(y_1)f(y_2)dy_1dy_2 = f(\rho\cos\varphi_1)f(\rho\sin\varphi_1)\rho d\rho d\varphi_1$$

and the joint density of random variables Z_1, Z_2 would be

$$\int_0^\infty f(\rho\cos\varphi_1)f(\rho\sin\varphi_1)\rho d\rho.$$

Let

$$\int_0^\infty f(\rho\cos\varphi_1)f(\rho\sin\varphi_1)\rho d\rho = c, \tag{2.10}$$

where c is some constant.

Now the question is: does the function $f(.)$ represent the density of the normal distribution? Substitution $r = \rho\cos\varphi_1$ in equation (2.10) gives

$$\int_0^\infty f(r)f(r\tan\varphi_1)rdr = c\cos^2\varphi_1, \tag{2.11}$$

or

$$\int_0^\infty f(r)f(tr)rdr = \frac{c}{1+t^2}, \tag{2.12}$$

where $t = \tan\varphi_1$. Multiply both sides of (2.12) by t^s and integrate with respect to t on the interval $(0, \infty)$:

$$\int_0^\infty \int_0^\infty f(r)f(tr)rdrt^sdt = \int_0^\infty \frac{c}{1+t^2}t^sdt. \tag{2.13}$$

The inner integral

$$\int_0^\infty f(tr)t^sdt = \frac{1}{r^{s+1}}\int_0^\infty f(y)y^sdy.$$

Then

$$\int_0^\infty f(r)\frac{1}{r^s}dr\int_0^\infty f(y)y^sdy = \int_0^\infty \frac{c}{1+t^2}t^sdt. \tag{2.14}$$

Expression

$$\varphi(s) = \int_0^\infty f(y)y^sdy \tag{2.15}$$

is the Mellin transform of the function $f(x)$. Then

$$\varphi(-s)\varphi(s) = h(s), \tag{2.16}$$

where

$$h(s) = c\int_0^\infty \frac{t^s}{1+t^2}dt.$$

But

$$\int_0^\infty \frac{x^s}{1+x^2}dx = \frac{1}{2}\sec\frac{\pi s}{2}.$$

For the normal density

$$\varphi(s) = c2^{\frac{s-1}{2}}\Gamma\left(\frac{1-s}{2}\right)$$

and we will have

$$\Gamma\left(\frac{1+s}{2}\right)\Gamma\left(\frac{1-s}{2}\right) = \left(\Gamma\left(\frac{1}{2}\right)\right)^2\frac{1}{\cos\frac{s\pi}{2}}$$

Thus, for the density $f(x)$, which is an even function we have

$$\varphi(s) = \int_0^\infty f(y)y^s dy$$

and the equation

$$\varphi(s)\varphi(-s) = c\frac{1}{\cos\frac{\pi s}{2}}.$$

From the identity

$$\Gamma\left(1-z\right)\Gamma\left(z\right) = \frac{\pi}{\sin\pi z}$$

by changing variables as $z = (1-s)/2$ we obtain

$$\Gamma\left(\frac{1+s}{2}\right)\Gamma\left(\frac{1-s}{2}\right) = \frac{\pi}{\sin\frac{\pi(1-s)}{2}}.$$

If we denote $\varphi(-s) = \psi(\frac{1-s}{2})$ and $\varphi(s) = \psi(\frac{1+s}{2})$, we have

$$\psi\left(\frac{1-s}{2}\right)\psi\left(\frac{1+s}{2}\right) = \frac{c}{\cos\frac{\pi s}{2}}$$

and finally

$$g(1-z)g(z) = \frac{\pi}{\sin\pi z},$$

where $z = (1-s)/2$. But it is not necessary (Laha, 1958) for $g(z)$ to be Gamma function. From

$$\cos x = \prod_{n=1}^\infty\left(1-\frac{x^2}{\pi^2(n-1/2)^2}\right)$$

we get

$$\cos\frac{\pi s}{2} = \prod_{n=1}^\infty\left(1-\frac{s^2}{(2n-1)^2}\right)$$

and

$$\frac{1}{\cos\frac{\pi s}{2}} = \prod_{n=1}^\infty\frac{1}{1-\frac{s^2}{(2n-1)^2}}.$$

and by different "regrouping" we get

$$\varphi(s)\varphi(-s) = \prod_{n=1}^{\infty} \frac{1}{1 - \frac{s^2}{(2n-1)^2}} \varphi^2(0),$$

keeping $\varphi(s)$ as a Mellin transform.

Thus, one can make the following statement.

Theorem 2.3.4. *For the case $k = 2$ the result obtained by Sakata is not true, that is, there are non-normal random variables for which $\mathbb{Z}$ is uniformly distributed over the sphere Φ_+^1.*

2.3.2.2 Case $k = 3$

In this case the characterization obtained by Sakata holds, and it allows one to test the normality of the gene expression data. Of course, in view of this characterization, it is enough to test the uniformity of the distribution of the vector $\mathbb{Z}$ on Φ_+^2. To do this, we use statistics, generated by $\mathfrak{N}$-distances, which will be discussed in the next chapter.

Chapter 3

Test of spherical uniformity

3.1 Introduction

Characterization theorems from chapter 2 allow one to replace normality test of the original sample with the test of uniformity on the sphere of the transformed sample. The transformed sample is just a measurable function of the observed random variables or some statistic, and hence the uniformity test on the sphere could be explained in terms of the distributions of such statistics.

Our approach for testing uniformity on the sphere is mainly based on the characterization theorems. The test statistic that we use in this chapter is called $\mathfrak{N}$-distance statistic for uniformity test on the hypersphere (Bakshaev, 2010) and is based on $\mathfrak{N}$-distances (Zinger et al., 1989).

3.2 Statistics for uniformity test on the sphere

Since our approach is based on statistics derived from $\mathfrak{N}$-distances, we will give a few necessary definitions and statements on $\mathfrak{N}$-distances from Klebanov (2005). We will use the same notations as in Klebanov (2005).

3.2.1 $\mathfrak{N}$-distances

Let $\mathfrak{X}$ be a non-empty set and $\mathcal{L} : \mathfrak{X}^2 \to \mathbb{C}$, where $\mathbb{C}$ is the complex plane.

Definition 3.2.1 (Negative definite kernel). *We shall say that $\mathcal{L}$ is a negative definite kernel if for any $n \in \mathbb{N}$, arbitrary points $x_1, \ldots, x_n \in \mathfrak{X}$ and any complex numbers $c_1, \cdots, c_n$ under the condition $\sum_{j=1}^{n} c_j = 0$ the following inequality holds:*

$$\sum_{j=1}^{n} \sum_{j=1}^{n} \mathcal{L}(x_i, x_j) \leq 0. \tag{3.1}$$

Definition 3.2.2 (Strictly negative definite kernel)**.** *We shall say that negative definite kernel $\mathcal{L}$ is strictly negative definite if the equality in (3.1) is true for $c_1 = \ldots = c_n = 0$ only.*

Definition 3.2.3 (Strongly negative definite kernel)**.** *Let* Q *be a measure on* $(\mathfrak{X}, \mathfrak{M})$ *and h be an integrable with respect to Q function such that*

$$\int_{\mathfrak{X}} h(x) dQ(x) = 0.$$

We shall say that $\mathcal{L}$ is strongly negative definite kernel if it is negative definite and the equality

$$\int_{\mathfrak{X}} \int_{\mathfrak{X}} \mathcal{L}(x, y) h(x) h(y) dQ(x) dQ(y) = 0$$

implies that $h(x) = 0$ Q-almost everywhere for any measure Q.

For the case $\mathfrak{X} = \mathrm{R}^d$ as examples of strongly negative definite kernels one can mention

$$\mathcal{L}(x, y) = \|x - y\|^{\alpha}, \quad 0 < \alpha < 2,$$

$$\mathcal{L}(x, y) = \frac{\|x - y\|}{1 + \|x - y\|},$$

$$\mathcal{L}(x, y) = \log(1 + \|x - y\|^2),$$

where $\|\cdot\|$ is the usual euclidean norm.

Let $(\mathfrak{X}, \mathfrak{M})$ be a measurable space and $\mathfrak{B}$ be the set of all probability measures on it. Suppose that $\mathcal{L}$ is a real continuous function and denote by $\mathfrak{B}_{\mathcal{L}}$ the set of all probability measures μ for which

$$\int_{\mathfrak{X}} \int_{\mathfrak{X}} \mathcal{L}(x, y) d\mu(x) d\mu(y) < +\infty.$$

Denote

$$\begin{aligned} \mathcal{N}(\mu, \nu) &= 2 \int_{\mathfrak{X}} \int_{\mathfrak{X}} \mathcal{L}(x, y) d\mu(x) d\nu(y) - \\ &- \int_{\mathfrak{X}} \int_{\mathfrak{X}} \mathcal{L}(x, y) d\mu(x) d\mu(y) - \int_{\mathfrak{X}} \int_{\mathfrak{X}} \mathcal{L}(x, y) d\nu(x) d\nu(y), \end{aligned}$$

where $\mu, \nu \in \mathfrak{B}_{\mathcal{L}}$.

Theorem 3.2.1. *(Klebanov, 2005) Let $\mathcal{L}$ be a real continuous function on $\mathfrak{X}^2$ under the condition*

$$\mathcal{L}(x, y) = \mathcal{L}(y, x), \quad x, y \in \mathfrak{X}. \tag{3.2}$$

Then the inequality

$$\mathcal{N}(\mu, \nu) \geq 0 \tag{3.3}$$

holds for all $\mu, \nu \in \mathfrak{B}_{\mathcal{L}}$ if and only if $\mathcal{L}$ is negative definite kernel. Inequality (3.3) holds with equality in the case $\mu = \nu$ only, if and only if $\mathcal{L}$ is strongly negative definite kernel.

Theorem 3.2.2. *(Klebanov, 2005) Let $\mathcal{L}$ be a real continuous function on $\mathfrak{X}^2$ satisfying*

$$\mathcal{L}(x,y) = \mathcal{L}(y,x), \quad \mathcal{L}(x,x) = 0, \quad \textit{for all} \quad x, y \in \mathfrak{X}. \tag{3.4}$$

Then

$$\mathfrak{N} = \mathcal{N}^{1/2}(\mu, \nu)$$

is a distance on $\mathfrak{B}_{\mathcal{L}}$.

3.2.2 Statistics based on $\mathfrak{N}$-distances

Let $X_1, \ldots, X_n$ be observations of the random variable X, where $X_i \in \mathbb{R}^d$ and $\|X_i\| = 1$, $i = 1, \ldots, n$. Suppose we test the hypothesis H_0 that X has a uniform distribution on the sphere Φ^{d-1}. Then for the kernel

$$\mathcal{L}(x,y) = \|x - y\|^\alpha, \quad 0 < \alpha < 2,$$

and $d = 3$ the statistic based on $\mathfrak{N}$-distance has a form

$$T_n = (2R)^\alpha \frac{2n}{\alpha + 2} - \frac{1}{n} \sum_{i,j=1}^{n} \|X_i - Xj\|^\alpha, \tag{3.5}$$

where R is radius of the sphere (Bakshaev, 2008).

The asymptotic distribution of T_n is defined as

$$\frac{3}{4} T_n \to \sum_{k=1}^{\infty} a_k^2 \chi^2_{2k+1}, \tag{3.6}$$

where χ^2_{2k+1} are independent chi-square random variables with $2k+1$ degrees of freedom and

$$a_k^2 = \frac{1}{2} \int_0^\pi (1 - \frac{3}{2} \sin \frac{x}{2}) \sin x P_k(\cos x) dx, \tag{3.7}$$

$P_k(x)$ are Legendre polynomials.

Note that for $d = 2$ statistics of the type (3.5) and their asymptotic distributions similar to (3.6) are also derived by Bakshaev (Bakshaev, 2008). But due to characterization theorems which require a minimal sample size of 3, they cannot be used for the purposes of testing spherical uniformity and hence normality. This is from one side. From the other side, in principle, one can consider the uniformity test on the hypersphere of any dimension greater than 3. But statistics of the form (3.5) are not available for $d > 3$. Therefore, for spherical testing the best available option would be $d = 3$. For the case $d = 1$, related results were published by Shokirov (Shokirov, 2007).

Results of the spherical uniformity test, which are related to this chapter were published in Shokirov (2013a). Some other results which are applicable to gene expression data are published in Shokirov (2012, 2013b). They are not included in this thesis.

3.3 Application of the $\mathfrak{N}$-distance based test to gene expression data

3.3.1 Calculating statistic T_n and its p values

For testing uniformity on the sphere we use the kernel

$$\mathscr{L} = \|x - y\| .$$

Then for $d = 3$ and $R = 1$ from (3.5) by using Sakata's transformations (2.7) we obtain

$$T_n = n - \frac{3}{2n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|A_i - A_j\|, \tag{3.8}$$

where

$$A_i = \left\{ \frac{z_{3i-2}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}}, \frac{z_{3i-1}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}}, \frac{z_{3i}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}} \right\}$$

denotes the i-th raw of the matrix $A = (a_{ij})$, and

$$a_{ij} = \frac{z_{3(i-1)+j}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}}, \ i = 1, \dots, n, \ j = 1, 2, 3.$$

Since

$$\begin{aligned}
\|A_i - A_j\|^2 &= \left(\frac{z_{3i-2}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}} - \frac{z_{3j-2}}{\sqrt{z_{3j-2}^2 + z_{3j-1}^2 + z_{3j}^2}} \right)^2 + \\
&+ \left(\frac{z_{3i-1}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}} - \frac{z_{3j-1}}{\sqrt{z_{3j-2}^2 + z_{3j}^2 + z_{3j}^2}} \right)^2 + \\
&+ \left(\frac{z_{3i}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2}} - \frac{z_{3j}}{\sqrt{z_{3j-2}^2 + z_{3j-1}^2 + z_{3j}^2}} \right)^2 = \\
&= 2 \left(1 - \frac{z_{3i-2} z_{3j-2} + z_{3i-1} z_{3j-1} + z_{3i} z_{3j}}{\sqrt{z_{3i-2}^2 + z_{3i-1}^2 + z_{3i}^2} \sqrt{z_{3j-2}^2 + z_{3j-1}^2 + z_{3j}^2}} \right) = \\
&= 2 \left[1 - (a_{i1} a_{j1} + a_{i2} a_{j2} + a_{i3} a_{j3}) \right],
\end{aligned}$$

then the statistic T_n could expressed in the form

$$T_n = n - \frac{3}{\sqrt{2} n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sqrt{1 - (a_{i1} a_{j1} + a_{i2} a_{j2} + a_{i3} a_{j3})}. \tag{3.9}$$

Calculating T_n by (3.9) is more efficient than by (3.8). Therefore, when testing uniformity we take advantage of it in calculations.

To calculate p-values of T_n, first by equation (3.7) we obtain a series of coefficients a_k^2, $k = 1, \ldots$. Since in (3.6) χ^2_{2k+1} are independent chi-square random variables with the characteristic function

$$f(t) = (1 - 2it)^{-\frac{2k+1}{2}},$$

then for some finite n and every a_k^2, $k = 1, 2, \ldots, n$ from (3.6) we obtain

$$f_n(t) = \prod_{k=1}^{n} (1 - 2ia_k^2 t)^{-\frac{2k+1}{2}}.$$

Then distribution of T_n would be

$$F_T(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_n(t) \frac{e^{-itx}}{it} dt.$$

From this equation for large values of T_n we obtain p-values as

$$p = 1 - F_T(x) = \mathrm{P}\left\{T_n > x\right\}. \tag{3.10}$$

3.3.2 Uniformity test of the gene expression data on the sphere

Let m be the number of the gene expression levels and n the sample size for each gene expression. Then the matrix $\mathbb{X} = (X_{ij}),\ i = 1, \ldots, n, j = 1, \ldots, m$ denotes n observations of m gene expression levels; element X_{ij} denotes the i-th observed level of the j-th gene. We proceed with the spherical uniformity test in the following way:

- Order columns of the matrix $\mathbb{X}$ by increasing variance;
- From matrix $\mathbb{X}$ construct matrix $\mathbb{Y} = (Y_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, k,\ k = [m/2]$, where $Y_{ij} = X_{i2j} - X_{i,2j-1}$ (δ-sequences);
- Construct matrix $\mathbb{Z}$, where $Z_{ij} = Y_{2i-1j} - Y_{2i,j}$, $i = 1, \ldots, s$, $s = n/2$, $j = 1, \ldots, k$ (symmetrization). If n is an odd number then skip the last observation;
- Construct matrix $\mathbb{U}_{l \times k}$ with dimensions $l \times k$, where $l = 3[s/3]$ in the following way. Here and afterward, unless otherwise specified, $[a]$ denotes the integer part of the number a.
  - For $i = 1, \ldots, s$, take the i-th column of the matrix $\mathbb{Z}$ and split its elements by groups of three elements;
  - Normalize each group of three elements (Sakata's transformations);
  - Merge all groups of three elements and save it as the i-th row of the matrix $\mathbb{U}$;
- Merge all columns of the matrix $\mathbb{U}$ to obtain a sample of size $l \times k$;

- Split obtained in the previous step sample of the size $l \times k$ by groups of three elements;
- Calculate T statistics as in (3.9) and its p-values according to (3.10) to verify uniformity on sphere Φ_+^2.

3.3.2.1 Results by testing uniformity on the sphere

Table 3.1: Results of spherical uniformity test

Size of random sample	p-value of the T statistic
100	0.565657
200	0.358586
300	0.526936
400	0.850168
500	0.250842
1000	0.181818
2000	0.111111
5000	0.076835
10000	0.011121

As seen from the Table 3.1, p-values calculated by the random samples of the sizes less than 5000 are relatively high. Therefore, we cannot reject uniformity of the transformed, hence the normality of the original data. The more we increase the size of the randomly chosen sample, the smaller p-values we obtain. For samples of size greater than 10000 p-values are very small (close to zero). This means that for samples of sizes greater than 10000 we reject the normality. In any case, we cannot make a definite statement regarding rejection or acceptance of the hypothesis that gene expression data are normally distributed.

3.3.3 Modifications of the uniformity test on the sphere

If random variables X and Y follow normal distribution with mean μ and the variance σ^2, then the difference $X - Y$ has normal distribution with zero mean and the variance $2\sigma^2$. Due to this fact we can, to some extent, modify the above-mentioned testing procedure for spherical uniformity.

When testing the uniformity of the gene expression data on the sphere we proceed with the assumption that the original data are from normal distribution, without specifying parameters, and due to transformations by Sakata they should follow standard normal distribution. As we mentioned above, we do not test normality of the gene expression data themselves but the differences of their two neighborhood expression levels, that is, the δ-sequences, calculated after ordering gene expressions by increasing variance. Here also we deal with δ-sequences.

Thus, if X_{ij}, the i-th observed level of the j-th gene ($i = 1, \dots, n, j = 1, \dots, m$), is an observation that follows the normal law with the mean μ and the variance σ^2, then the

δ-sequences $Y_{ij} = X_{2ij} - X_{2i-1,j}$, $i = 1, \ldots, s$, $(s = [n/2])$, $j = 1, \ldots, m$ should follow normal with zero mean and the variance $2\sigma^2$. A non-standard normal distribution (including one with zero mean) could be transformed into standard normal. Therefore, we proceed with the test of normality (or the same uniformity test on the sphere) by transformation of the original gene expression data into assumed standard normal. Below, we give an algorithm for this type of testing.

3.3.3.1 Algorithm 1

Here and below we consider each column of the matrix $\mathbb{X}$ as a vector (of dimension 88).

1. Order columns of the matrix $\mathbb{X}$ by increasing variance;
2. Take $2k$ vectors (corresponding to $2k$ columns of the matrix $\mathbb{X}$) $\boldsymbol{X}_j = \{X_{1j}, \ldots X_{nj}\}$, $j = 1, \ldots, 2k$, where $1 \le k \le 3542 = [m/2]$ and $n = 88$.
3. Construct k vectors $\boldsymbol{Y}_j = \{Y_{1j}, \ldots, Y_{nj}\}$, where $Y_{ij} = X_{i,2j} - X_{i,2j-1}$, $i = 1, \ldots, n$, $j = 1, \ldots, k$;
4. Construct k vectors $\boldsymbol{Z}_j = \{Z_{1j}, \ldots, Z_{sj}\}$, where $Z_{ij} = Y_{2i,j} - Y_{2i-1,j}$, $i = 1, \ldots, s$, $s = [n/2]$, $j = 1, \ldots, k$;
5. Estimate the covariance matrix

$$\Sigma = \begin{pmatrix} S_{z_1z_1} & S_{z_1z_2} & \ldots & S_{z_1z_s} \\ S_{z_2z_1} & S_{z_2z_2} & \ldots & S_{y_2z_s} \\ \ldots & \ldots & \ldots & \ldots \\ S_{z_sz_1} & S_{z_sz_2} & \ldots & S_{z_sz_s} \end{pmatrix},$$

   where $S_{z_iz_j} = S_{z_jz_i}$ and

$$S_{z_iz_j} = \frac{1}{s-1}\sum_{i=1}^{s}\left[\left(Z_i - \frac{1}{s}\sum_{i=1}^{s} Z_i\right)\left(Z_j - \frac{1}{s}\sum_{j=1}^{s} Z_j\right)\right].$$

6. Take the l-th coordinate of vectors $\boldsymbol{Z}_j, j = 1, \ldots, k$, construct k-dimensional vectors $\{Z_{l1}, \ldots, Z_{lk}\}$ for all $l = 1, \ldots, s$ and transform them into k-dimensional vectors $\{U_{l1}, \ldots, U_{lk}\}$, $l = 1, \ldots, s$ by

$$\{U_{l1}, \ldots, U_{lk}\} = \Sigma^{-1/2}\{Z_{l1}, \ldots, Z_{lk}\}, \quad l = 1, \ldots, s.$$

7. From the first, second, etc. k-th coordinates of the vectors $\{U_{l1}, \ldots, U_{lk}\}$, $l = 1, \ldots, s$ construct s-dimensional vectors $\boldsymbol{U}_j = \{U_{1j}, \ldots, U_{sj}\}, j = 1, \ldots, k$.
8. Merging vectors $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k$ obtain vector $\boldsymbol{V} = \{V_1, \ldots, V_{sk}\}$ of dimension $s \times k$, where $V_j = U_{1j}, V_{k+j} = U_{2j}, V_{2s+j} = U_{3j}, \ldots, V_{(s-1)k+j} = U_{sj}$ for $j = 1, \ldots, k$.

9. Split vector $\boldsymbol{V}$ by groups of three elements, normalize them: divide each group by its norm and obtain the matrix $A = (a_{ij})_{M\times 3}$, where

$$a_{ij} = \frac{V_{3(i-1)+j}}{\sqrt{(V_{3i-2})^2 + (V_{3i-1})^2 + (V_{3i})^2}}, \; i = 1, \ldots, M = [sk/3], j = 1, 2, 3.$$

10. Calculate T statistics by (3.9) and its p-values according to (3.10) to verify uniformity on sphere Φ_+^2 and normality of the original data.

The results of the test by this Algorithm are shown in the Table 3.2.

3.3.3.2 Results of uniformity test by Algorithm 1

Table 3.2: The results of uniformity test by Algorithm 1

Sample size (k)	p-value of the T-statistic
102 (7)	0.498316
205 (14)	0.525253
308 (21)	0.469697
498 (34)	0.621212
601 (41)	0.052187
997 (68)	0.005051
2009 (137)	0.001818
5001 (231)	0.000168

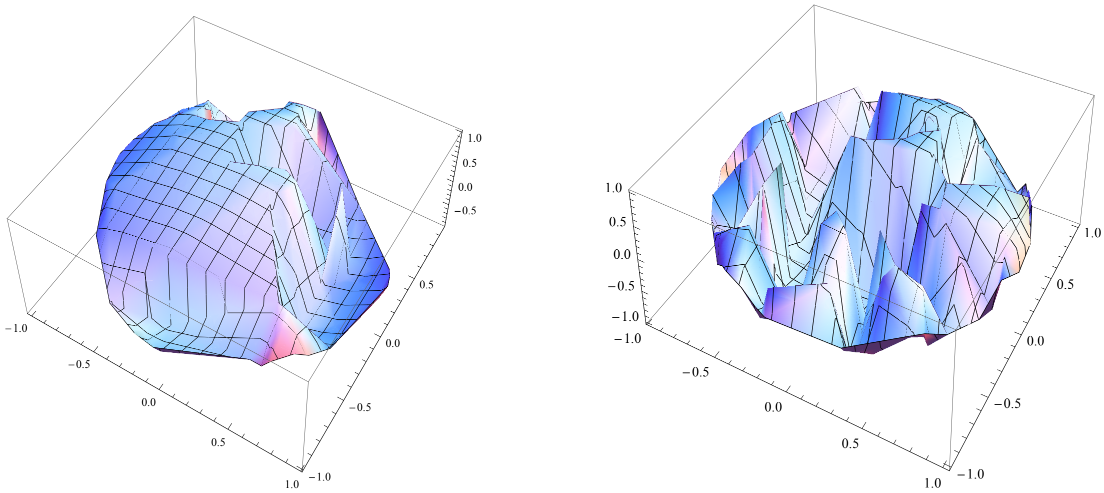
As seen from Table 3.2, p-values obtained by Algorithm 1 have some similarities and some dissimilarities with those in the previous testing procedure: they are relatively large for small samples and smaller for larger samples. But in contrast to previous testing procedure we do not observe a decreasing pattern of p-values for samples of small sizes. Even so they are smaller and smaller as the sample size increases. So as in the case of the previous testing procedure, we cannot reject uniformity transformed data on the sphere for smaller samples and hence the normality of the original data. For large samples we reject uniformity on the sphere, hence the normality of the δ-sequences of the gene expression data.

Remark 3.3.1. *1. This testing procedure allows one to conduct a uniformity test for any* $2k$ *(randomly) chosen number of gene expressions from* $k = 1$ *to* $k = 3542$*;* $k = 1$ *corresponds to 2 columns of the matrix* $\mathbb{X}$ *(or 2 gene expressions and* $k =$ 3542 *corresponds to complete number of genes expressions, 7084.*

2. If we randomly choose $2k$*-vectors (*$k = 1, 2, \ldots, 3542$*), then we have samples of sizes* $14, 29, 44, \ldots, 51949$ *of the transformed spherical data. Due to taking integer parts of the numbers of columns and rows of the matrix* $\mathbb{X}$*, these numbers do not exactly match the corresponding sample sizes in the previous testing procedure.*

3.3.3.3 Observations on the the sphere

Figure 3.1 shows a visualization of observations on the three-dimensional sphere. Each observation is a vector of unit length.



(A) Sample size 2000

(B) Sample size 1000

Figure 3.1: This is a visualization of a subset of the transformed observations on the unit sphere. Each observation is a three-dimensional vector $u = (x, y, z)$, which corresponds to a point on the sphere. These observations are obtained by Sakata transformations from HYPERDIP data-set according to the testing procedure for normality test. 3.1 (a) consists of 2000 three-dimensional observations and 3.1(b) of 1000 such observations.

Chapter 4

One-dimensional test and the distribution of its statistic

4.1 Introduction

Following the principle of the reconstruction for the complete type of distributions by the studentized differences, in this chapter we prove a characterization theorem for the normal distribution. Based on that, we use Kolmogorov's statistic to test the normality of the gene expression data.

We extend the characterization theorem of the normal distribution for the reconstruction of the complete type of distributions by the distribution of a nonlinear statistic. Based on this theorem, we test whether the two samples of the gene expression data belong to the same type of distributions. For this purpose, we apply Kolmogorov-Smirnov's two-sample test statistic.

4.2 The U test

Let $X_1, \ldots, X_n$ be a sample of size $n \geq 9$. Construct the following linear forms:

$$L = X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3, \tag{4.1}$$

$$L_0 = X_4 - \frac{1}{2}X_5 - \frac{1}{2}X_6, \tag{4.2}$$

$$L_1 = X_7 - \frac{1}{2}X_8 - \frac{1}{2}X_9. \tag{4.3}$$

It is clear that if the random variables $X_1, \dots, X_n$ have normal distribution with the mean μ and variance σ^2, then the linear forms $L,\ L_0$ and L_1 will have normal distribution with zero mean and the variance $\frac{3}{2}\sigma^2$.

Consider the statistic

$$U = \frac{L}{\sqrt{|L_0 L_1|}}. \tag{4.4}$$

The distribution of the statistic U is expressed in terms of the Meijer G-function, which will be defined below.

4.2.1 The Meijer G-function

The G-function was introduced by C. S. Meijer (Meijer, 1936) as a very general function intended to include most of the known special functions as particular cases. The first definition was given by using a series (see also Meijer (1940, 1941a,b). Today's definition, which is more general, is represented via a line (curve or contour) integral in the complex plane, introduced in its full generality by Arthur Erdélyi in 1953 (Bateman and Erdélyi, 1953). This or definition of the Meijer G-function can be found in many sources, for example, Mathai and Saxena (1973); Prudnikov et al. (1990); Gradshteyn and Ryzhik (2000, 2007).

According to Erdélyi the Meijer G-function is defined as

$$\begin{aligned} G^{m,n}_{p,q}\left(z \,\middle|\, \begin{matrix} \boldsymbol{a} \\ \boldsymbol{b} \end{matrix}\right) &= G^{m,n}_{p,q}\left(z \,\middle|\, \begin{matrix} a_1, \dots, a_p \\ b_1, \dots, b_q \end{matrix}\right) = \\ &= \frac{1}{2\pi i}\int_\gamma \frac{\prod_{j=1}^m \Gamma(b_j - s)\prod_{j=1}^n \Gamma(1 - a_j + s)}{\prod_{j=m+1}^q \Gamma(a_j - s)\prod_{j=n+1}^p \Gamma(1 - b_j + s)} z^s ds \end{aligned} \tag{4.5}$$

where the empty product is interpreted as 1, $0 \le m \le q$, $0 \le n \le p$, m, n, p and q are integer numbers, and the parameters are such that no pole of any $\Gamma(b_j - s),\ j = 1, \dots m$ coincides with any pole of any $\Gamma(1 - a_k + s),\ k = 1, \dots, n$; this assumption could be written as $a_k - b_j \ne 1, 2, 3, \dots$ for $k = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$, and

$$\Gamma(t) = \int_0^\infty z^{t-1} e^{-z} dz$$

is the Euler's Gamma function.

Three types of integration paths γ in the right member of (4.5) can be exhibited:

1. The path γ runs from $-i\infty$ to $+i\infty$ such that all poles of the functions $\Gamma(b_j - s),\ j = 1, 2, \dots, m$, lie to the right, and all poles of the functions $\Gamma(1 - a_k + s),\ k = 1, 2, \dots, n$ lie to the left of the path γ. In this case the conditions under which the integral on the right side of equation (4.5) converges are of the form

$$p + q < 2(m + n), \quad |\arg(z)| < (m + n - \frac{1}{2}p - \frac{1}{2}q)\pi.$$

2. The path γ is a loop, beginning and ending at $+\infty$ that encircles the poles of the functions $\Gamma(b_j - s)$ (for $j = 1, 2, \dots, m$) once in the negative direction. All the poles of the functions $\Gamma(1 - a_k + s)$ (for $k = 1, 2, \dots, n$) must stay outside the loop. Then, under which the integral on the right side of equation (4.5) converges are:

$$q \geq 1 \quad \text{and either} \quad p < q \text{ or} \quad p = q \quad \text{and} \quad |z| < 1.$$

3. The path γ is a loop, beginning and ending at $-\infty$, that encircles the poles of the functions $\Gamma(1 - a_k + s)$ for $k = 1, 2, \dots, n$) once in the positive direction. All the poles of the functions $\Gamma(b_j - s)$ for $j = 1, 2, \dots, m$ must remain outside the loop. The conditions under which the integral (4.5) converges are

$$p \geq 1 \quad \text{either} \quad p > q \quad \text{or} \quad p = q \quad \text{and} \quad |z| > 1.$$

It is assumed that the values of the parameters and the variable z are such that at least one of these three definitions makes sense. In cases where more than one of the definitions make sense, they lead to the same result.

The function $G^{m,n}_{p,q}\left(z \,\middle|\, \begin{matrix} \boldsymbol{a} \\ \boldsymbol{b} \end{matrix}\right)$ is analytic with respect to z; it is symmetric with respect to the parameters $a_1, \dots, a_n$ and also with respect to $a_{n+1}, \dots, a_p$, $b_1, \dots, b_m$ and $b_{m+1}, \dots, b_p$.

One of the properties of the G-function that we use later is the following: The G-function with $p > q$ can be transformed into the G-function with $p < q$ by means of the relationships:

$$G^{m,n}_{p,q}\left(z^{-1} \,\middle|\, \begin{matrix} a_1, \dots, a_p \\ b_1, \dots, b_q \end{matrix}\right) = G^{m,n}_{p,q}\left(z \,\middle|\, \begin{matrix} 1 - a_1, \dots, 1 - a_p \\ 1 - b_1, \dots, 1 - b_q \end{matrix}\right) \tag{4.6}$$

4.2.2 A characterization theorem of the normal distribution

By taking logarithm of both sides of the equation (4.4) we obtain

$$L_y = Y_1 - \frac{1}{2}Y_2 - \frac{1}{2}Y_3, \tag{4.7}$$

where $L_y = \ln|U|$, $Y_1 = \ln|L|$, $Y_2 = \ln|L_0|$ and $Y_3 = \ln|L_1|$. L_y is a linear form, similar to L and its distribution determines the distribution of Y_1 to within a location parameter if it has an analytic ch.f. on some stripe around real axis. We show that it also determines the distribution of X_1 to within a scale parameter.

To prove the characterization of the normal density, we need the following statement.

Lemma 4.2.1. *Characteristic functions of the random variables Y_1, Y_2 and Y_3 are analytic functions of the real variable t.*

Proof. Since Y_1, Y_2 and Y_3 are identically distributed, it is enough to show that ch. f. of the random variable Y_1 is an analytic function. Let us denote by $f_{Y_1}(t)$ the ch. f. of the r. v. Y_1. Then

$$\begin{aligned} f_{Y_1}(t) &= \mathbb{E}\left[e^{itY_1}\right] = \mathbb{E}\left[e^{it\ln|L|}\right] = \\ &= \mathbb{E}\left[|L|^{it}\right] = \int_{-\infty}^{\infty} |x|^{it}\varphi_L(x)dx, \end{aligned}$$

where

$$\varphi_L(x) = \frac{1}{\sigma\sqrt{3\pi}}e^{-\frac{x^2}{3\sigma^2}}$$

is the density function of the r. v. L.

Since $\varphi_L(x)$ is an even function, then

$$f_{Y_1}(t) = 2\int_0^{\infty} x^{it}\varphi_L(x)dx = \frac{2}{\sigma\sqrt{3\pi}}\int_0^{\infty} x^{it}e^{-\frac{x^2}{3\sigma^2}}dx. \tag{4.8}$$

It is well-known (see, for example, Gradshteyn and Ryzhik (2007, 1130-1132)) that for $s \in \mathbb{C}$ such that $\Re(s) > 0$ the following holds

$$\int_0^{\infty} x^{s-1}e^{-x^2}dx = \frac{1}{2}\Gamma\left(\frac{s}{2}\right). \tag{4.9}$$

Equation (4.9) is the Mellin transform of the function e^{-x^2}, the normal density, where

$$\Gamma(s) = \int_0^{\infty} x^{s-1}e^{-x}dx$$

denotes the Euler's Gamma function.

By using equation (4.9) from (4.8) we obtain

$$f_{Y_1}(t) = \frac{(\sigma\sqrt{3})^{it}}{\sqrt{\pi}}\Gamma\left(\frac{1+it}{2}\right). \tag{4.10}$$

Since $f_{Y_1}(t)$ is expressed in terms of gamma-function and gamma-function is meromorphic, hence $f_{Y_1}(t)$ is an analytic function of t in some strip around real axis. □

Theorem 4.2.1 (A characterization theorem of the normal density)**.** *Let $X_1, \ldots, X_n$ ($n \geq 9$) be a sample from a population with the normal density with mean μ and the variance σ^2. Consider the linear forms*

$$L = X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3 \tag{4.11}$$

and

$$L_y = Y_1 - \frac{1}{2}Y_2 - \frac{1}{2}Y_3, \tag{4.12}$$

where $L_y = \ln|U|$, $Y_1 = \ln|L|$, $Y_2 = \ln|L_0|$ and $Y_3 = \ln|L_1|$. Then the distribution of the linear forms L and L_y determine the normal distribution to within a location and a scale parameter, respectively.

Proof. We first show that the distribution of the linear form L determines the distribution of X_1 to within a location parameter. This means that if

$$\widetilde{L} = \widetilde{X}_1 - \frac{1}{2}\widetilde{X}_2 - \frac{1}{2}\widetilde{X}_3,$$

where $\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3$ are independent and identically distributed random variables, is another linear form which is identically distributed with the linear form L, then its distribution differs from the distribution of L to within a location parameter. Since L is distributed normally, then $\widetilde{L}$ should have a normal distribution as well. Hence, as it follows from Cramer's theorem (Cramér, 1936), random variables $\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3$ have normal distribution too. Let random variables $\widetilde{X}_1$, $\widetilde{X}_2$ and $\widetilde{X}_3$ have normal distribution with the mean $\widetilde{\mu}$ and variance $\widetilde{\sigma}^2$. Then the linear form $\widetilde{L}$ has the normal distribution with zero mean and the variance $3\widetilde{\sigma}^2/2$.

If we denote by $f_L(t)$ and $f_{\widetilde{L}}(t)$ the characteristic functions of the linear forms L and $\widetilde{L}$, respectively, then

$$f_L(t) = f(t)f^2\left(-\frac{t}{2}\right) = \exp\left(-\frac{3\sigma^2}{4}t^2\right)$$

and

$$f_{\widetilde{L}}(t) = \widetilde{f}(t)\widetilde{f}^2\left(-\frac{t}{2}\right) = \exp\left(-\frac{3\widetilde{\sigma}^2}{4}t^2\right),$$

where $f(t)$ and $\widetilde{f}(t)$ are characteristic functions of the random variables X_1 and $\widetilde{X}_1$, respectively. Since both linear forms have normal distribution with zero mean, then from the last two equations it follows that $\sigma = \widetilde{\sigma}$.

Now we show that distribution of the linear form L_y determines distribution of the r. v. X_1 to within a scale parameter. As before, let

$$L_z = Z_1 - \frac{1}{2}Z_2 - \frac{1}{2}Z_3, \tag{4.13}$$

where Z_1, Z_2 and Z_3 are independent and identically distributed random variables, be another linear form, identically distributed with the linear form L_y. Since distributions of the linear forms L_y and L_z are different from normal distribution, we cannot apply Cramer's theorem as we did in the case with the linear form L.

Let $f(t)$ be a ch. f. of the r. v. Y_1 and $g(t)$ be a ch. f. of the r. v. Z_1. Then the ch. f. of the linear forms L_y and L_z are

$$f_{L_y}(t) = f(t)f^2(-t/2)$$

and

$$f_{L_z}(t) = g(t)g^2(-t/2),$$

respectively, and the condition of identically distributed linear forms L_y and L_z is

$$f(t)f^2(-t/2) = g(t)g^2(-t/2). \tag{4.14}$$

Characteristic function $f_{L_y}(t)$ is a real analytic function and, from Raikov's theorem on analytic characteristic functions (see, for example, Ramachandran (1967); Lukacs (1970); Linnik and Ostrovskii (1977)), it follows that its components, $f(t)$ and $f(-t/2)$, are analytic functions as well. From equation (4.14) it follows that the ch. f. $g(t)$ should have the same strip of analyticity as $f(t)$. Since $f(t)$ is a ch.f., then $f(0) = 1$. From this follows that $f(t) \neq 0$ for $|t| < \delta$ for some $\delta > 0$. Then we can consider $\log f(t)$ for $|t| < \delta$. Since $\log f(t)$ is a multivalued function, we consider the branch where $f(t)$ is different from zero. The same can be said about the ch.f. $g(t)$, that is, $g(t) \neq 0$ for $|t| < \rho$ for some $\rho > 0$ and we consider that branch of $\log g(t)$ where $g(t) \neq 0$.

Then from equation (4.14) we have

$$\varphi(t) + 2\varphi(-t/2) = \psi(t) + 2\psi(-t/2), \quad \varphi(0) = \psi(0) = 0. \tag{4.15}$$

where $\varphi(t) = \log f(t)$ and $\psi(t) = \log g(t)$, for $|t| < \min\{\delta,\ \rho\}$.

Since characteristic functions $f(t)$ and $g(t)$ are expressed in terms of gamma-function and gamma-function is a meromorphic function, then $f(t)$ and $g(t)$ are analytic functions, and as it follows from Lemma 4.2.1 they have finite derivatives of all orders at the point $t = 0$. By taking derivatives of the k-th ($k = 1, 2, \ldots$) order from both sides of the equation (4.15) at $t = 0$, we have

$$\left(1 + 2(-1)^k \frac{1}{2^k}\right) \varphi^{(k)}(0) = \left(1 + 2(-1)^k \frac{1}{2^k}\right) \psi^{(k)}(0). \tag{4.16}$$

From (4.16) it follows that for $\{t : |t| < \min\{\delta,\ \rho\}\}$ and $k = 1$ $\varphi'(0)$ and $\psi'(0)$ can be any numbers but for $k > 1$

$$\varphi^{(k)}(0) = \psi^{(k)}(0).$$

In particular, $k = 2$ we have

$$\varphi''(0) = \psi''(0), \quad |t| < \min\{\delta,\ \rho\}. \tag{4.17}$$

Form equation (4.17) it follows that

$$\psi(t) = bt + \varphi(t)$$

for some constant b and

$$g(t) = e^{bt} f(t) \tag{4.18}$$

for real t.

For the characteristic functions $f(t)$ and $g(t)$ we have $|f(t)| \leq 1$ and $|g(t)| \leq 1$. Since

$$g(-t) = \overline{g(t)}$$

and

$$f(-t) = \overline{f(t)},$$

then from the equation (4.18) it follows that $b = i\theta$, where $\theta \in R^1$ and i is the imaginary unit. So we have

$$g(t) = e^{i\theta t} f(t). \tag{4.19}$$

Equation (4.19) shows that the d.f. of the r.v. Y_1 differs from the d.f. of the r.v. Z_1 only to within a location parameter and the logarithm shows that it differs from the distribution of the r v. X_1 only to within a scale parameter.

□

Theorem 4.2.2. *Let $X_1, \ldots, X_n$, $n \geq 9$ be a sample from normal distribution with the mean μ and the variance σ^2 and let r. v. U be defined as in (4.4). Then U has the d. f.*

$$F_U(x) = \frac{1}{2} + \frac{1}{2^{3/2}\pi^4 x^2} G^{5,\,4}_{5,\,5}\left(\frac{4}{x^4} \,\middle|\, \begin{matrix} -\frac{1}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2} \\ -\frac{1}{2}, -\frac{1}{4}, 0, 0, \frac{1}{4} \end{matrix} \right) \tag{4.20}$$

and the characteristic function

$$\varphi_U(t) = \frac{t^2}{16\pi^{7/2}} G^{6,2}_{2,6}\left(\frac{t^4}{64} \,\middle|\, \begin{matrix} -\frac{1}{4},\ \frac{1}{4} \\ -\frac{1}{2},\ -\frac{1}{4},\ 0,\ 0,\ 0,\ \frac{1}{4} \end{matrix} \right), \tag{4.21}$$

where

$$G^{m,n}_{p,q}\left(x \,\middle|\, \begin{matrix} \boldsymbol{a} \\ \boldsymbol{b} \end{matrix} \right) = G^{m,n}_{p,q}\left(x \,\middle|\, \begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix} \right)$$

is the Meijer G-function.

Proof. As before, let $f(t)$ be a ch. f. of the r. v. Y_1. Then the ch. f. $f_{L_y}(t)$ of the linear form L_y is

$$f_{L_y}(t) = f(t) f^2(-t/2).$$

As it was shown in the proof of Lemma 4.2.1, the characteristic function of the random variables Y_1 is

$$f(t) = \frac{\left(\sqrt{3}\sigma\right)^{it}}{\sqrt{\pi}} \Gamma\left(\frac{1+it}{2}\right). \tag{4.22}$$

Then

$$f_{L_y}(t) = \frac{1}{\pi^{3/2}} \Gamma\left(\frac{1+it}{2}\right) \left[\Gamma\left(\frac{2-it}{4}\right)\right]^2 \tag{4.23}$$

is the ch.f. of the liner form L_y.

By the inverse Fourier transform from (4.23) we obtain the density function of L_y:

$$\varphi_{L_y}(x) = \frac{1}{2\pi} \frac{1}{\pi^{3/2}} \int_{-\infty}^{\infty} e^{-itx} \Gamma\left(\frac{1+it}{2}\right) \left[\Gamma\left(\frac{2-it}{4}\right)\right]^2 dt. \tag{4.24}$$

By change of variable $s = it$ we have

$$\varphi_{L_y}(x) = \frac{1}{2\pi i}\frac{1}{\pi^{3/2}}\int_{-i\infty}^{i\infty} e^{-sx}\Gamma\left(\frac{1+s}{2}\right)\left[\Gamma\left(\frac{2-s}{4}\right)\right]^2 ds. \tag{4.25}$$

Since $L_y = \ln|U|$, then the density of the r. v. U would be

$$\varphi_U(x) = \frac{1}{2\pi i}\frac{1}{\pi^{3/2}}\int_{-i\infty}^{i\infty} x^{-(s+1)}\Gamma\left(\frac{1+s}{2}\right)\left[\Gamma\left(\frac{2-s}{4}\right)\right]^2 ds, \tag{4.26}$$

or by replacing $s+1$ to s,

$$\varphi_U(x) = \frac{1}{2\pi i}\frac{1}{\pi^{3/2}}\int_{-i\infty}^{i\infty} x^{-s}\Gamma\left(\frac{s}{2}\right)\left[\Gamma\left(\frac{3-s}{4}\right)\right]^2 ds, \tag{4.27}$$

which represents the inverse Mellin transform of the moment generating function of the r. v. L_y. By the change of variable from s to $4s$ in (4.27) and using the property of the gamma-function,

$$\Gamma(nx) = (2\pi)^{\frac{1-n}{2}} n^{nx-\frac{1}{2}} \prod_{k=0}^{n-1}\Gamma\left(x+\frac{k}{n}\right),$$

after some manipulations, we obtain

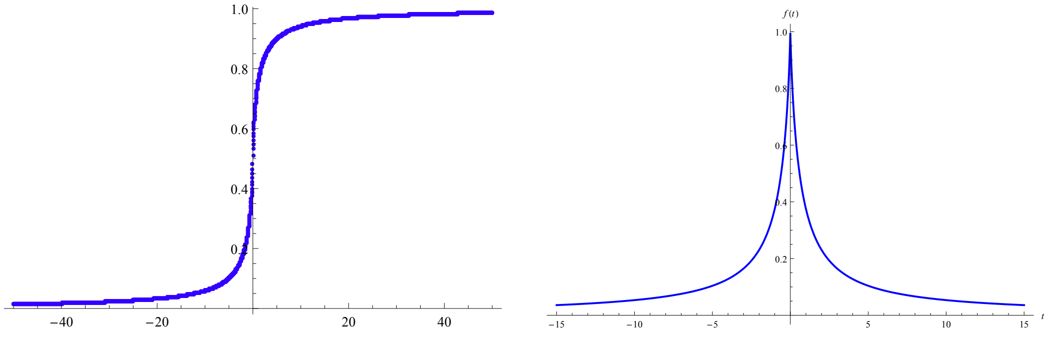
$$\varphi_U(x) = \frac{1}{2\pi i}\frac{2}{\pi^{3/2}}\int_{-i\infty}^{i\infty}\left(\frac{4}{x^4}\right)^{-s}\Gamma(s)\Gamma\left(s+\frac{1}{2}\right)\left[\Gamma\left(\frac{3}{4}-s\right)\right]^2 ds. \tag{4.28}$$

Equation (4.28) represents the density function of the r.v. U and it is the Mellin transform, which could be represented as the Meijer G-function. From this equation by calculating $\mathbb{E}\left[e^{itU}\right]$ and using the property (4.6) of the G function, after some more calculations, we obtain equation (4.21). By integrating equation (4.28) from $-\infty$ to x with respect to x and again using (4.6), we obtain (4.20). □

Remark 4.2.1. *In practice, the explicit representation of the d.f. of the statistic U is not as important as the reconstruction of the distribution of the original sample $X_1, \ldots, X_n$ by the distribution of the statistic U. When it comes to testing normality, instead of using the theoretical distribution of U one can use its empirical counterpart, constructed by a uniformly generated random sample, the size of which is comparable to the sample of the real data under the test.*

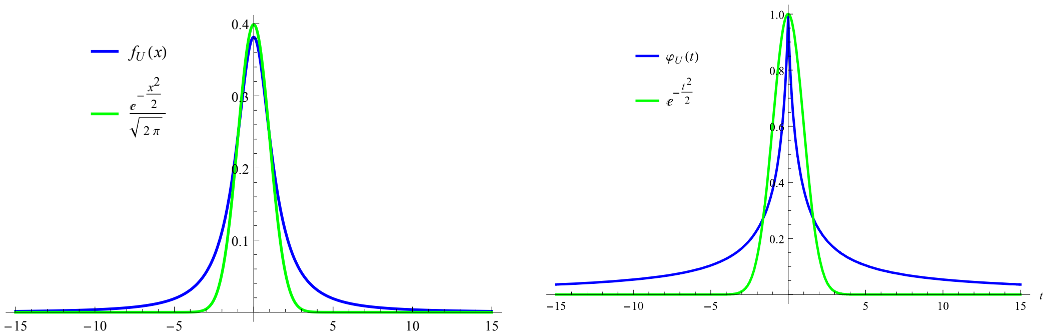
4.2.2.1 Graphs of the distribution, probability density and the characteristic functions of the statistic U

Distribution, density and the characteristic functions of the random variable (statistic) U are visualized in Figures 4.1 and 4.2. Probability density and the characteristic functions are shown in comparison with the density and the characteristic functions of the standard normal distribution. As shown in Figure 4.2, density and characteristic functions of the random variable U are similar to the density and characteristic functions of the standard normal distribution.



(A) Distribution function (B) Characteristic function

FIGURE 4.1: Distribution function and the characteristic function of the random variable (statistic) U. 4.1a is the cumulative distribution function of the random variable (statistic) U. 4.1b is the characteristics function of the random variable (statistic) U.



(A) Probability density function (B) Characteristic function

FIGURE 4.2: Probability density and the characteristic function of the statistic U in comparison with those of the standard normal distribution. Figure 4.2a shows the probability density functions of the random variable (statistic) U (blue line) and the standard normal distribution (green line). Figure 4.2b shows the characteristic functions of the random variable (statistic) U (blue line) and the standard normal distribution (green line).

4.2.3 A characterization theorem of the complete type

Theorem 4.2.1 can be extended for the reconstruction of the complete type. As seen, the proof of Theorem 4.2.1 makes use of the analytic property of the characteristic function of normal distribution. In general, a distribution of the complete type, which is different from normal, may not have an analytic ch.f. Therefore, for the reconstruction of the complete type, assumption of having an analytic ch.f. is essential. Under validity of this assumption, the whole complete type can be reconstructed.

Let $X_1, \ldots, X_n$ be a sample of size n, $n \geq 9$. Construct the following linear forms:

$$L_x = X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3,$$

$$L_{x0} = X_4 - \frac{1}{2}X_5 - \frac{1}{2}X_6,$$

$$L_{x1} = X_7 - \frac{1}{2}X_8 - \frac{1}{2}X_9.$$

Consider the statistic

$$U = \frac{L_x}{\sqrt{|L_{x0}L_{x1}|}}. \tag{4.29}$$

The following statement holds.

Theorem 4.2.3. *Let $X_1, \ldots, X_n$, $(n \geq 9)$ be a sample from a distribution function $F(x, \theta)$, $\theta = (\mu, \sigma)$, where $\mu \in R$ and $\sigma > 0$ are location and scale parameters, respectively. Consider the linear forms*

$$L_x = X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3 \tag{4.30}$$

and

$$L_u = U_1 - \frac{1}{2}U_2 - \frac{1}{2}U_3, \tag{4.31}$$

where $L_u = \ln|U|$, $U_1 = \ln|L_x|$, $U_2 = \ln|L_{x0}|$, $U_3 = \ln|L_{x1}|$ and L_x, L_{x0} and L_{x1} are defined above. Assume that the random variables X_1 and U_1 have characteristic functions, analytic in some strip around the real axis. Then distribution of the linear forms L_x and L_u determine the distribution of $F(x)$ to within a location and scale parameters, respectively. In other words, if $\widetilde{X}_1, \ldots, \widetilde{X}_n$, $(n \geq 9)$ is a sample from a distribution $G(x)$ and the linear form

$$\widetilde{L}_x = \widetilde{X}_1 - \frac{1}{2}\widetilde{X}_2 - \frac{1}{2}\widetilde{X}_3 \tag{4.32}$$

is identically distributed with the linear form L_x and

$$\widetilde{L}_u = \widetilde{U}_1 - \frac{1}{2}\widetilde{U}_2 - \frac{1}{2}\widetilde{U}_3, \tag{4.33}$$

is another linear form, identically distributed with L_u then

$$F(x, \theta) = G\left(\frac{x - \mu}{\sigma}\right).$$

Proof. Proof follows from theorem 4.2.1. □

From the Theorem 4.2.3 it follows that if $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two samples of sizes n and m $(m, n > 9)$ then by the distribution of the linear forms

$$L_u = U_1 - \frac{1}{2}U_2 - \frac{1}{2}U_3, \tag{4.34}$$

and

$$L_v = V_1 - \frac{1}{2}V_2 - \frac{1}{2}V_3, \tag{4.35}$$

where

$$L_u = \ln|U|, \quad U_1 = \ln|L_x|, \quad U_2 = \ln|L_{x0}|, \quad U_3 = \ln|L_{x1}|$$

and

$$L_V = \ln|V|, \quad V_1 = \ln|L_y|, \quad V_2 = \ln|L_{y0}|, \quad V_3 = \ln|L_{y1}|,$$

one can reconstruct their distributions. Hence by using Kolmogorov-Smirnov's statistic we can test if both of samples are from the same distribution or belong to the same type. Since Kolmogorov-Smirnov's test statistic makes use of the empirical d.f.'s, there is no need for the explicit representation of distribution of the statistic L_U and L_V.

4.3 Application of the Kolmogorov and Kolmogorov - Smirnov's statistics to the gene expression data

In this section we describe the procedure of the normality test of the gene expression data and test of whether two samples of the gene expressions belong to the same type.

To test normality we use Kolmogorov's statistics, defined as

$$D = \sup_x |F_n(x) - F(x)|, \tag{4.36}$$

where $F_n(x)$ is the empirical distribution function. Since Kolmogorov's statistic is a distribution free statistic, then instead of (4.36) we can use the statistic

$$D = \sup_x |U_n(x) - U(x)|, \tag{4.37}$$

where $U(x)$ denotes the uniform distribution function and $U_n(x)$ is the empirical distribution function of the uniform sample $U_1, \ldots, U_n$ with $U_i = F(X_i)$, $i = 1, \ldots, n$.

To test whether two samples of the gene expressions data belong to the same type, we use Kolmogorov-Smirnov's two- sample test, defined as

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|,$$

where $F_m(x)$ and $G_n(x)$ are empirical distribution functions, constructed by the corresponding samples.

4.3.0.1 The procedure of normality test

Let m be the number of the gene expression levels and n sample size for each gene expression. Then the matrix $\mathbb{X} = (X_{ij})$, $i = 1, \ldots, n, j = 1, \ldots, n$ denotes n observations of m gene expression levels; element X_{ij} denotes the i-th observed level of the j-th gene; rows of the matrix $\mathbb{X}$ correspond to observations and columns to the genes.

The procedure for the normality test is as follows:

1. From matrix $\mathbb{X}$ construct a matrix $\mathbb{Y} = (Y_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, k,\ k = [m/2]$, where $Y_{ij} = X_{i2j} - X_{i,2j-1}$ (δ-sequences);

2. For $j = 1, \ldots, k$, take the j-th column of the matrix $\mathbb{Y}$ and separate it by groups of 9 elements (and again inside by (3) groups of 3 elements); the number of groups with 9 elements would be $l = [n/9]$.

3. Compute linear forms $L,\ L_0,\ L_1$ as in equations (4.1) and (4.3) and "observation" (statistic) U as in the equations (4.4) and obtain a matrix of observations $\mathbb{Z} = (Z_{ij}),\ i = 1, \ldots, l, j = 1, \ldots, k$;

4. Merge all columns of $\mathbb{Z}$ and obtain a sample of size kl.

5. From the sample $\mathbb{Z}$ randomly choose a sample of size s, $s = 10,\ 20,\ 100,\ 1000, \ldots, kl$ and apply the d. f. of U to obtain a uniform random sample.

6. Test uniformity (normality) of the obtained random sample from the previous step by using Kolmogorov's statistics.

4.3.1 Results of the normality test

We used the above testing procedure to test normality of the gene expressions in the data set HYPERDIP, consisting of 88 observations of 7084 gene expression levels. By randomly chosen samples, we calculated Kolomogorov's statistic and its p-values. Repeating this for each sample size 102 times, we took the average of the value of Kolmogorov's statistic and its corresponding p-values. Obtained results (mean values of the value of Kolmogorov's statistic and its corresponding p-values) are shown below in Table 4.1.

Table 4.1: Results of normality test

Size of random sample	Kolmogorov's statistic	
	Mean value	Mean p-value
100	0.0971641	0.2829700
200	0.0723744	0.2340880
300	0.0677369	0.1217220
400	0.0598150	0.1097140
500	0.0616188	0.0430078
1000	0.0594583	0.0026269
2000	0.0537717	0.0000180
5000	0.0571029	1.4×10^{-14}
10000	0.0577657	0.000000

Table 4.1 shows that the average of the p-values, calculated by random samples of sizes less than 1000 are relatively high. Therefore, we do not have enough evidence to reject normality, for example at the significant level 0.001. For the samples of larger sizes it seems to be less likely for the data to follow the normal distribution. Therefore we cannot make a strong statement whether the gene expression levels in the HYPERDIP data set follow normal distribution or not.

4.3.1.1 Testing procedure of whether the two samples belong to the same type

For this test, we basically repeat the procedure of the normality test twice, once with each sample or data set. Let matrix $\mathbb{X} = (X_{ij})$, $i = 1, \ldots, m, j = 1, \ldots, p$ and matrix $\mathbb{Y} = (Y_{ij})$, $i = 1, \ldots, n, j = 1, \ldots, q$, where m and n denote the samples sizes (number of rows of the matrices $\mathbb{X}$ and $\mathbb{Y}$), p and q be the number of gene expressions. Since we have the same number of genes in both data sets, we assume that $p = q$.

Testing procedure is conducted as follows:

1. For the matrix $\mathbb{X}$ repeat steps of the normality testing procedure from step 1 to 5. As the result of obtain a (one-dimensional) sample $\mathbb{U}$ of size M.

2. For the matrix $\mathbb{Y}$ repeat steps of the normality testing procedure from step 1 to 5 and obtain the sample $\mathbb{V}$ of size N (numbers M and N can be calculated precisely through m, n and p)

3. From the samples $\mathbb{U}$ and $\mathbb{V}$ randomly choose a sample of size s, $s = 10$, 20, 100, $1000, \ldots, M(N)$ and construct corresponding empirical distributions.

4. Calculate p-value of Kolmogorov-Smirnov's statistic and decide whether or not to reject the hypothesis that the two samples $\mathbb{U}$ and $\mathbb{V}$ are from the same type of distributions.

4.3.2 Test results of whether the two samples belong to the same type

We tested whether or not the two data-sets of gene expressions belong to the same type by the above-mentioned procedure. As the matrix $\mathbb{X}$ is taken the HYPERDIP data set that consists of 88 observations and as the matrix $\mathbb{Y}$ is taken TEL data set that consists of 79 observations. In both data sets each observation has the same dimensions of 7084. As before, by randomly chosen samples, according to the above-mentioned testing procedure, we calculated Kolmogorov-Smirnov's statistic and its corresponding p-values. We repeated this procedure many times, and calculated the mean value of Kolmogorov-Smirnov's statistic and the mean of its p-values. Corresponding results are shown below in Table 4.2.

As Table 4.2 shows, for samples of sizes up to 500 the the average of p-value of Kolmogorov-Smirnov's statistic is relatively large, so we cannot reject the hypothesis that data of the gene expression levels in the data sets HYPERDIP and TEL belong to the same type of distributions (for example, at the significant level 0.001). But for the samples of larger sizes (greater then 500) we reject the hypothesis. If we calculate Kolmogorov-Smirnov's statistic for the complete samples $\mathbb{U}$ and $\mathbb{V}$, then we obtain a p value very close to zero. Since, for the small samples, we cannot reject the hypothesis of belonging to the same type and for the larger samples we can, just as we did for the normality test, we cannot make a strong statement as to whether these two data sets belong to the same type of

Table 4.2: The test results of whether the two samples belong to the same type

Size of random sample	Kolmogorov-Smirnov's statistic	
	Mean value	Mean p-value
100	0.1000	0.702057
200	0.0700	0.712339
300	0.1133	0.042339
400	0.1075	0.019656
500	0.1100	0.004716
1000	0.0890	0.000726
2000	0.0715	0.000073
5000	0.0740	2.6×10^{-9}

distribution or not. We can say that there are genes in these data sets which have distributions of the same type (larger p-values for small samples), but for a general statement we do not have enough evidence. This situation can be explained, in particular, by the fact that these two data sets may not contain data of all (expressed) genes.

Conclusion

There were many efforts in statistical literature to verify the validity of the assumption on normality of the gene expression data, some evidencing this assumption, while some others not. This persuades us one to further investigate this question.

Statistical analysis of the available data can be performed in different ways. In particular, testing whether this data set follows some probability law, for example the normal distribution. When testing normality, one can just apply some statistics, for example, Shapiro-Wilks test of (non) normality and make a conclusion on hypothesis under the test. This can be considered as one of the approaches. But before proceeding with any test some knowledge on the structure and nature of the data should be clarified.

In the case of gene expression data, which was considered in this thesis, before discussing the normality test, we first gave a general overview of how these data are derived. Starting with microarray experiment, we concluded with the explanation of the correlation structure of the gene expressions and highlighted some viewpoints on utilizing the dependence between gene expressions.

As shown, one of the main issues is the dependence between gene expressions data, complicated correlations of the pair-wise gene structure. Another issue is the small number of observations. If there exist certain ways to benefit from the dependence of the gene expressions, there is no way to increase the number of observations. By using some technique that reduces the dependence, we were able to take advantage of the small sample sizes. Once we consider the expressed level of all genes as independent multidimensional observations, then independence allows one to consider each gene separately. Then, pooling all genes together yields a very large one-dimensional sample of the gene expressions. Namely, this was the main principle of transforming gene expression data into δ-sequences.

Mainly this work was devoted to studying normality of the gene expressions data. In comparison with the traditional testing procedures, we applied a different approach. For testing normality, we proceeded from the idea of reconstructing the type of distributions. After giving the main results on the reconstruction of the type of distributions, considering normal distribution as a complete type with location and scale parameters, we transformed normality test to the spherical uniformity test.

Spherical uniformity, and hence the normality test of the gene expression data, was conducted by the statistics based on $\mathfrak{N}$-distances.

In this thesis we proved a new result on reconstruction of the complete type of distributions. In particular, a theorem for the characterization of the normal density was given.

The task of verifying normality of the gene expressions was conducted by applying the existing results of reconstruction, uniformity test and the characterization theorems for the complete type of distributions, in particular for the normal density.

The results of the uniformity and normality tests, represented in Chapters 3 and 4, in some cases give evidence on normality of the gene expression and some other cases against it. As seen from the results of the normality test, there are more evidences against normality than there are in favor of normality. As previously explained, to test normality we basically transformed the whole gene expression data into one huge sample. Then, we randomly chose samples of different sizes. For samples of relatively small sizes we could not reject normality but when we increased the sample size, the result was that we obtained very small p-values, so there were enough evidences for larger samples to reject the hypothesis of normality, which was against our expectations, since by the law of large numbers one can expect convergence to the normal distribution.

When analyzing the results of the normality test (Table 4.1) and the test of whether the two samples belong to the same type (Table 4.2), we observe the following situation. With the test of normality we cannot reject normality for sample size less or equal to 400. We reject the hypothesis of normality of the gene expressions data for the samples of sizes less than or equal to 500 at the significant level 0.05 or for the samples of sizes greater than or equal to 1000 at the significant level 0.001. But with the test of whether the two samples belong to the same type we reject the hypothesis of whether the two data sets follow the same type of distributions, at the significant level 0.05 for the samples of sizes less than or equal to 300 already. This situation can be explained by the influence of the subset of modified genes or variations of the genes in both data sets. In both data sets the majority of the gene expressions may follow the normal distribution but because of the biological changes due to the illness, the subset of the modified genes in the first data set does not coincide with those in the second one. This causes a larger difference between the distributions of the gene expression levels in the two data sets, hence leads to the rejection of the hypothesis of whether the two samples belong to the same type for the samples of relatively small sizes than those in the normality test. Based on this, we propose biological explanation of the observed fact. The gene expressions corresponding to the illness have non-normal distribution, while the others are normally distributed.

As for the statistical verification of the normality assumption of the gene expression data, we cannot reject the hypothesis on normality assumption for a large (but not too large) number of observations. So we cannot make a firm definitive statement either in favor or against this assumption. Therefore, the question posed at the beginning of this thesis remains open: the assumption whether or not the gene expressions are distributed according to the Normal Distribution needs more exploration.

Theoretical results obtained in this thesis can be applied to other multidimensional observations too, for example, the data of the Taiwanese–American Occultation Survey

System (TAOS) (Meinshausen and Rice, 2006). Besides being applicable to practical problems, these results are of theoretical interest, too.

The list of author's publications

Publications related to the thesis

1. Bobosharif K. Shokirov, *On characterization of the complete type of distributions* (submitted).

2. Bobosharif Shokirov, *Test for normality of the gene expression data.* In the book *"Statistical Methods for Microarray Data Analysis: Methods and protocols"*, Yakovlev, A. Y, Klebanov, L. B. and Gaile, D. (Eds.), Springer, 2013, pp. 193-208. DOI:10.1007/978-1-60327-337-412.

3. Bobosharif K. Shokirov, *A lower bound for the mixture parameter in the binary mixture model and its estimator. In Antoch, J. et al. (Eds.) Proceedings of the 17th Summer School of the Union of Czech Mathematicians and Physicists*, pp. 117-125, 12/2013.

4. Bobosharif K. Shokirov, *On a problem connected with the mixture parameter estimation. In Antoch, J. et al. (Eds.) Proceedings of the 16th Summer School of the Union of Czech Mathematicians and Physicists*, **Vol. 3**, pp. $95-102$, 2010.

5. Bobosharif K. Shokirov, *Two-sample test by using $\mathfrak{N}$-distances.* WDS'07 Proceedings of Contributed papers: Part I - Mathematics and Computer Sciences, pp. 193-197, 2007.

Other publication

1. Bobosharif K. Shokirov, *O resheniaych odnorodnoy obobshyonnoy systemy Cauchy-Riemann.* Dokladi AN Tadzikistan (in Russian), Dushanbe, 2000. Eng. trans: On the solutions of homogeneous system of the Cauchy-Riemann.

2. Shokirov B.K., Baizaev A., *On an initial value problem, connected with the generalized Cauchy-Riemann system* (in Russian), Khujand, 2002.

3. Kabilov, M. M., Shokirov B. K,. Satriddinov, P. B., *On spatial stability stationary regime of filtration of gases combustion* (in Russian), Dushanbe, 2007, Reports of the Academy of Sciences of the Republic of Tajikistan.

4. Bobosharif K. Shokirov, *On the solutions of generalized Cauchy–Riemann system*, *Complex Variables and Elliptic Equations*, 59(8), pp.1118-1130, 2014. doi: 10.1080/17476933.2013.816843.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2015). *Molecular Biology of The Cell.* Garland Science, NewYork, Sixth edition.

Ali, A. (2014). Biological aspects: Types and applications of microarrays. In Rueda, L. (Ed.), *Microarray Image and Data Analysis: Theory and Practice*, Chapter 2. CRC Press (Taylor and Francis), New York.

Altman, R. and Raychaudhuri, S. (2001). Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol.*, 11:340–347.

Amaratunga, D. and Cabrera, J. (2001a). Analysis of data from viral dna microchips. *Journal of the American Statistical Association*, 96(456):1161–1170.

Amaratunga, D. and Cabrera, J. (2001b). Outlier Resistance, Standardization, and Modeling Issues for DNA Microarray Data. In Fernholz, L., Morgenthaler, S., and Stahel, W. (Eds.), *Statistics in Genetics and in the Environmental Sciences*, Trends in Mathematics, pages 17–26. Birkhäuser Basel.

Amaratunga, D., Cabrera, J., and Shkedy, Z. (2014). *Exploration and Analysis of DNA Microarray and Other High-Dimensional Data.* Wiley Series in Probability and Statistics. Wiley, The second of 2003 book by Amaratunga, Dhammika and Cabrera, Javier "Exploration and Analysis of DNA Microarray and Protein Array Data" edition.

Arteaga-Salas, J. M., Zuzan, H., Langdon, W. B., Upton, G. J. G., and Harrison, A. P. (2008). An overview of image-processing methods for Affymetrix genechips. *Briefings in Bioinformatics*, 9(1):25–33.

Babu, M. M. (2004a). Introduction to microarray data analysis. *Computational Genomics: Theory and Application*, pages 225–249.

Babu, M. M. (2004b). Introduction to microarray data analysis. In Grant, P. R. (Ed.), *Computational Genomics: Theory and Application*, Chapter 1, pages x + 306. Horizon Bioscience, Laboratory of Molecular Biology, Cambridge, UK.

Bakshaev, A. (2008). Nonparametric tests based on $\mathfrak{N}$-distance. *Lithuanian Mathematical Journal*, 48(4):368–379.

Bakshaev, A. (2010). $\mathfrak{N}$-distance tests of uniformity on the hypersphere. *Nonlinear Analysis: Modeling and Control*, 15(1):15–28.

Bateman, H. and Erdélyi, A. (1953). *Higher Transcendental Functions*, Volume I. McGraw–Hill, New York.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bolsover, S. R., Shephard, E. A., White, H., and S., H. J. (2011). *Cell Biology: A short Course*. John Wiley and Sons, Inc., Hoboken, New Jersey, 3rd edition.

Chen, L., Klebanov, L., and Yakovlev, A. (2007). Normality of gene expression revisited. *Journal of Biological Systems*, 15(01):39–48.

Chu, T., Glymour, C., Scheines, R., and Spirtes, P. (2003). A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, 19(9):1147–1152.

Colantuoni, C., Henry, G., Bouton, C., Zeger, S., and Pevsner, J. (2003). Snomad: Biologist-friendly web tools for the standardization and normalization of microarray data. In Parmigiani, G., Garrett, E., Irizarry, R., and Zeger, S. (Eds.), *The Analysis of Gene Expression Data*, Statistics for Biology and Health, pages 210–228. Springer New York.

Cramér, H. (1936). Über eine eigenschaft der normalen verteilungsfunktion. *Mathematische Zeitschrift*, 41(1):405–414.

Draghici, S. (2012). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. Chapman and Hall, CRC Press, Boca Raton.

Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science*, 251(4995):767–773.

Galambos, J. and Kotz, S. (1978). *Characterizations of Probability Distributions*, Volume 657 of *Lecture Notes in Mathematics*. Springer, Berlin Heidelberg New York Tokyo.

Giles, P. J. and Kipling, D. (2003). Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 19(17):2254–2262.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Gradshteyn, I. S. and Ryzhik, I. M. (2000). Special functions. In Zwillinger, A. J. (Ed.), *Table of Integrals, Series, and Products (Sixth Edition)*, pages 851 – 1038. Academic Press, San Diego, sixth edition edition.

Gradshteyn, I. S. and Ryzhik, I. M. (2007). *Table of integrals, series, and products*. Elsevier/Academic Press, Amsterdam, seventh edition. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

Grant, G., Manduchi, E., and Stoeckert, Christian, J. (2002). Using non-parametric methods in the context of multiple testing to determine differentially expressed genes. In Lin, S. and Johnson, K. (Eds.), *Methods of Microarray Data Analysis*, pages 37–55. Springer US.

Graves, P. R. and Haystead, T. A. J. (2002). Molecular biologist's guide to proteomics. *Microbiology and Molecular Biology Reviews*, 66(1):39–63.

Guan, Z. and Zhao, H. (2005). A semiparametric approach for marker gene selection based on gene expression data. *Bioinformatics*, 21(4):529–536.

Hardin, J. and Wilson, J. (2009). A note on oligonucleotide expression values not being normally distributed. *Biostatistics*, 10(3):446–450.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J., Snesrud, E., Lee, N., and Quackenbush, J. (2000). A concise guide to cdna microarray analysis. *Biotechniques*, 29(3):548–563.

Hu, R., Qiu, X., Glazko, G., Klebanov, L., and Yakovlev, A. (2009). Detecting intergene correlation changes in microarray analysis: a new approach to gene selection. *Bioinformatics*, 10(20):446–450.

Istepanian, R. S. (2003). Microarray image processing: Current status and future directions. *IEEE Transactions on Nanobioscience*, 2(4):173 – 175.

Kagan, A. M., Linnik, Yu. V., and Rao, C. R. (1973). *Characterization problems in mathematical statistics. Translated from Russian text by B. Ramachandran.* Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley & Sons, a Wiley-Interscience Publication. XII, 499 p. Translation of "*Характеризационные задачи математической статистики.*" (Kharakterizatsionnye zadachi matematicheskoi statistiki).

Kagan, A. M. and Zinger, A. A. (1977). On the problem of reconstruction of the type of a distribution. *Theory of Probability and Its Applications*, 21(2):389–392.

Kakosyan, A., Klebanov, L., and Melamed, A. (1984). *Characterization of Distributions by the Method of Intensively Monotone Operators*, Volume 1088 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, Berlin Heidelberg New York Tokyo.

Klebanov, L., Gordon, A., Xiao, Y., Land, H., and Yakovlev, A. (2006a). A permutation test motivated by microarray data analysis. *Computational Statistics & Data Analysis*, 50(12):3619–3628.

Klebanov, L., Jordan, C., and Yakovlev, A. (2006b). A new type of stochastic dependence revealed in gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 5(1):1 – 23.

Klebanov, L. and Yakovlev, A. (2006). Treating expression levels of different genes as a sample in microarray data analysis: Is it worth a risk?. *Statistical Applications in Genetics and Molecular Biology*, 5(1):1 – 11.

Klebanov, L. and Yakovlev, A. (2008). A nitty-gritty aspect of correlation and network inference from gene expression data. *Biology Direct*, 3(1):35.

Klebanov, L. B. (2005). $\mathfrak{N}$-*Distance and its applications.* The Karolinum Press, Prague.

Klebanov, L. B. and Yakovlev, A. A. (2007). Diverse correlation structures in gene expression data and their utility in improving statistical inference. *The Annals of Applied Statistics*, 1(2):538–559.

Kovalenko, I. N. (1958). On the reconstruction of the additive type of distributions by a sequence of independent observations. *Trans. of the All-Union Con. on Prob. Theory and Math. Statist. (Erevan, 1958), Izdatelstvo Akademii Nauk Arm. SSR, Erevan, 1960. (In Russian.)*, 1(2).

Laha, R. G. (1958). An example of a non-normal distribution where the quotient follows the cauchy law. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):222–223.

Lee, M.-L. T., Whitmore, G. A., Björkbacka, H., and Freeman, M. W. (2005a). Generalized rank tests for replicated microarray data. *Generalized Rank Tests for Replicated Microarray Data*, 4(1):1544–6115. PMID: 16078385.

Lee, M.-L. T., Whitmore, G. A., Björkbacka, H., and Freeman, M. W. (2005b). Nonparametric methods for microarray data based on exchangeability and borrowed power. *Journal of Biopharmaceutical Statistics*, 15(5):783–797. PMID: 16078385.

Li, M. M., Ankita Patel, M., and Hu, X. (2012). Clinical applications of microarrays in cancer in modern clinical molecular techniques. In Hu, P., Hegde, M., and Lennon, P. A. (Eds.), *Modern Clinical Molecular Techniques.* Springer, New York.

Linnik, Yu. V. (1956). To the question of determination of parent distribution by distribution of statistic. *Teor. Verojatnost. i Primenen.*, 1(4):466–476. in Russian.

Linnik, Yu. V. (1968). *Statistical Problems with Nuisance Parameters.* Translations of mathematical monographs. American Mathematical Society.

Linnik, Yu. V. and Ostrovskii, I. V. (1977). *Decomposition of random variables and vectors.* Providence, Rhode Islands: American Mathematical Society. Translated from Russian.

Lockhart, D., Brown, E., Wong, G., Chee, M., and Gingeras, T. (2000). Expression monitoring by hybridization to high density oligonucleotide arrays. US Patent 6,040,138.

Lodish, H., Berk, A., Kaiser, C. A., and Krieger, M. (2007). *Molecular Cell Biology.* W.H. Freemann and Co, New York, sixth edition.

Lukacs, E. (1970). *Characteristic functions.* New York: Griffins.

Marko, N. F. and Weil, R. J. (2012). Non-gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. *PLoS ONE*, 7(10):e46935.

Mathai, A. M. and Saxena, R. K. (1973). *Generalized Hypergeometric Functions with Applications in Statistics and Physical Sciences.* Berlin: Springer, sixth edition.

Meier-Ewert, S., Lange, J., Gerst, H., Herwig, R., Schmitt, A., Freund, J., Elge, T., Lehrach, H., Mott, R., and Herrmann, B. (1998). Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Research*, 26(9):2216–2223.

Meijer, C. S. (1936). Über whittakersche bzw. besselsche funktionen und deren produkte. *Nieuw Arch. Wiskd., II. Ser*, 18(4):10–39.

Meijer, C. S. (1940). Über eine erweiterung der laplace-transformation. I, II. *Proc. Akad. Wet. Amsterdam*, 43:599–608, 702–711.

Meijer, C. S. (1941a). Eine neue erweiterung der laplace-transformation. I, II. *Proc. Akad. Wet. Amsterdam*, 44:727–737, 831–839.

Meijer, C. S. (1941b). Multiplikationstheoreme für die funktion $G^{m,n}_{p,q}(z)$. *Proc. Akad. Wet. Amsterdam*, 44:1062–1070.

Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34(1):373–393.

Moore, J. A. (1972). *Heredity and Development.* Oxford University Press, New York, 2nd edition.

Nettleton, D., Hwang, J., Caldo, R., and Wise, R. (2006). Estimating the number of true null hypotheses from a histogram of p values. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):337–356.

Parmigiani, G., Garett, E., Irizarry, R., and Zeger, S. (2006). *The Analysis of Gene Expression Data: Methods and Software.* Statistics for Biology and Health. Springer New York.

Petty, R. D., Kerr, K. M., Murray, G. I., Nicolson, M. C., Rooney, P. H., Bissett, D., and Collie-Duguid, E. S. (2006). Tumor transcriptome reveals the predictive and prognostic impact of lysosomal protease inhibitors in non–small-cell lung cancer. *Journal of Clinical Oncology*, 24(11):1729–1744.

Piras, V. and Selvarajoo, K. (2015). The reduction of gene expression variability from single cells to populations follows simple statistical laws. *Genomics*, 105(3):137 – 144.

Prokhorov, Yu. V. (1965). On a characterization of a class of probability distributions by those of some statistics. *Teor. Veroyatnost. i Primenen.*, 10(3):479–487.

Prudnikov, A. P., Brychkov, Yu. A., and Marichev, O. I. (1990). *Integrals and Series: More Special Functions, Vol. 3.* Gordon and Breach Science Publishers, New York. Translated from the Russian by G. G. Gould. Table erratum Math. Comp. v. 65 (1996), no. 215, p. 1384.

Qiu, X., Brooks, A. I., Klebanov, L. B., and Yakovlev, A. A. (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, 6:120 – 11.

Ramachandran, B. B. (1967). *Advanced theory of characteristic functions.* Calcutta: Statistical Pub. Society.

Rueda, L. (2014). *Microarray Image and Data Analysis: Theory and Practice.* CRC Press (Taylor and Francis), New York.

Rueda, L. and Ali, A. (2014). Introduction to microarrays. In Rueda, L. (Ed.), *Microarray Image and Data Analysis: Theory and Practice*, Chapter 1. CRC Press (Taylor and Francis), New York.

Sakata, T. (1977a). A test of normality based o some characterization theorems. *Memoirs of the Faculty of Science, Kyushu University. Series A*, 31(2):221–225.

Sakata, T. (1977b). Two characterization theorems of normal density function. *Memoirs of the Faculty of Science, Kyushu University. Series A*, 31(2):215–219.

Schena, M. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):465–470.

Schena, M. (2002). *Microarray Analysis. Professional developer's guide series.* Hoboken, Wiley.

Schena, M. (2003). *Microarray Analysis.* Hoboken, Wiley-Liss.

Shokirov, B. (2013a). Test for normality of the gene expression data. In Yakovlev, A. Y., Klebanov, L., and Gaile, D. (Eds.), *Statistical Methods for Microarray Data Analysis*, Volume 972 of *Methods in Molecular Biology*, pages 193–208. Springer New York.

Shokirov, B. K. (2007). Two-sample testing by using $\mathfrak{N}$-distances. *WDS'07 Proceedings of Contributed papers: Part I - Mathematics and Computer Sciences*, 44(2):193–197.

Shokirov, B. K. (2012). On a problem connected with the mixture parameter estimation. In Antoch, J. and Dohnal, G. (Eds.), *Proceedings of the 16th Summer School of the Union of Czech Mathematicians and Physicists*, Number 4, pages 95–102, Praguue. Czech Statistical Society.

Shokirov, B. K. (2013b). A lower bound for the mixture parameter in the binary mixture model and its estimator. In Antoch, J. and Dohnal, G. (Eds.), *Proceedings of the 17th Summer School of the Union of Czech Mathematicians and Physicists*, Number 3, pages 117–125, Praguue. Czech Statistical Society.

Smith, M., Dunning, M., Tavare, S., and Lynch, A. (2010). Identification and correction of previously unreported spatial phenomena using raw illumina beadarray data. *BMC Bioinformatics*, 11(1):208.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98:10869–10874.

Speed, T. (2004). *Statistical analysis of gene expression microarray data.* CRC Press.

Stamey, T. A., Warrington, J. A., Caldwell, M. C., Chen, Z., Fan, Z., Mahadevappa, M., McNeal, J. E., Nolley, R., and Zhang, Z. (2001). Molecular genetic profiling of gleason grade 4/5 prostate cancers compared to benign prostatic hyperplasia. *The Journal of Urology*, 166(6):2171 – 2177.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(1):717–738.

Wyrick, J. and Young, R. (2003). Deciphering gene expression regulatory networks. *Current Opinion in Genetics and Development*, 12(1):130–136.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C.-H., Evans, W. E., Naeve, C., Wong, L., and Downing, J. R. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143.

Zembutsu, H., Ohnishi, Y., Tsunoda, T., Furukawa, Y., Katagiri, T., Ueyama, Y., Tamaoki, N., Nomura, T., Kitahara, O., Yanagawa, R., et al. (2002). Genome-wide cdna microarray screening to correlate gene expression profiles with sensitivity of 85 human cancer xenografts to anticancer drugs. *Cancer research*, 62(2):518–527.

Zinger, A. A. (1956). On a problem of A. N. Kolmogorov. *Vestnik Leningradskogo Universiteta*, 1(2):15–19.

Zinger, A. A. and Kagan, A. M. (1976). Sample mean as an estimator of location parameter in presence of a nuisance scale parameter. *Theory of Probability and Its Applications*, 35(3):15–28.

Zinger, A. A., Kakosyan, A. V., and Klebanov, L. B. (1989). Characterization of distributions by means of statistics and some probability metrics. *Stability of Stochastic Models*, pages 47–55. (in Russian).

Zinger, A. A. and Linnik, Yu. V. (1964). A characteristic property of the normal distribution. *Theory of Probability & Its Applications*, 9(4):624–626.

Index

δ-sequence, 20

Meijer G-function, 42

additive type, 24
additive type of distribution, 24
adenine, 10
Affymetrix, 6
analytic function, 44

cell, 5, 7, 9
chemical bonds, 10
Complementary DNA (cDNA), 9
complete type, 26

deoxyribonucleic acid, 1
Deoxyribonucleic acid (DNA), 8
differentiation, 13
DNA strand, 10
DNA transcription, 9
down-regulated, 1
driver gene(s), 11

enzyme, 9
Eukaryote, 20

Gamma function, 44
gene, 8
gene expression, 5
gene expressions, 1, 6
genetics, 1
genomics, 1
guanine, 10

heterologous expression, 9
high-density oligonucleotide, 15
High-density oligonucleotide arrays, 16
human health, 1
HYPERDIP, 53

inhibitor, 14
invariant statistic, 25
inverse Fourier transform, 47

Kolmogorov's statistic, 51
Kolmogorov-Smirnov's statistic, 41, 53

large number of small samples, 23
line integral, 42
linear form, 44, 50
linear statistic, 25

maximal invariant statistic, 25, 26
Mellin transform, 29, 44
meromorphic, 44
messenger RNA, 5
microarray technologies, 1
molecular biology, 1
mRNA, 5
multiplicative type, 25

near-independent, 20
normal distribution, 23
nucleic acids, 12
nucleotides, 10
nylon membrane arrays, 15
nylon membrane filters, 15

oligonucleotide, 6

protein degradation, 11
Proteins, 8

random variable (r.v.), 44
Real-Time Quantitative Reverse Transcription (qRT-PCR), 14
reconstruction the type of distributions, 23
reverse transcriptase, 9
Reverse Transcription PCR (RT-PCR), 14
Ribonucleic acid (RNA), 8
ribosomal RNA, 11
ribosome, 11
RNA transcripts, 1
RNA translation, 9
RNA-Seq, 7

single cells, 7
spherical uniformity, 7
spotted arrays, 15

Spotted cDNA Microarrays, 15
statistical analysis, 5
structure, yielding near-independent, 19
studentized differences, 24
sugarphosphate, 10

TEL, 53
transcriptome, 1
transcriptome-wide variability, 7
translation parameter, 24, 25
two-sample test, 41

uniform distribution, 27
up-regulated, 1

whole-genome sequence, 1