Charles University in Prague

Faculty of Mathematics and Physics

**DOCTORAL THESIS**

Mgr. Peter Bubelíny

# Multidimensional statistics and applications to study genes

Department of Probability and Mathematical Statistics

| | |
|---|---|
| Supervisor of the doctoral thesis: | prof. Lev Klebanov, DrSc. |
| Study programme: | Mathematics |
| Specialization: | Probability and Mathematical Statistics |

Prague 2014

**Acknowledgement**

I would like to thank to my parents and my aunt for being patient and supporting me during the whole study. Specially, I would like to thank to Prof. Lev Klebanov, DrSc. for valuable comments, remarks and overall help with this doctoral thesis.

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 ........... date ............. signature of the author

Název práce: Mnohorozměrná statistika a aplikace na studium genů

Autor: Mgr. Peter Bubelíny

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Microarrayová data genových expresí se skládají z několika tisíců genů a pouze několika desítek pozorování. Navíc, geny jsou mezi sebou silně závislé a data obsahují systematické chyby. Proto nám rozsah těchto dat nedovoluje rozumně odhadnout jejich korelační strukturu. U mnoha statistických problémů s mircoarrayovými daty musíme současně testovat tisíce hypotéz. Vzhledem k závislosti mezi geny, *p*-hodnoty těchto hypotéz jsou taky závislé. V této práci porovnáme běžné procedury mnohonásobného testování, které jsou vhodné pro závislé hypotézy. Běžný způsob, jak udělat microarrayová data méně závislá a částečně odstanit systematické chyby, je normalizovat je. Proto bylo navrhnuto několik nových normalizací a studovali jsme, jak různé normalizace ovlivňují testování hypotéz. Navíc jsme porovnali testy pro nalezení odlišně expresovaných genů nebo genových množin a nalezli několik zajímavých vlastností testů jako například strannost dvoj-výběrového Kolmogorov-Smirnovova testu a zajímavé chování Hotellingova testu pro závislé složky pozorování. Na konci jsme navrhli test pro testování nezávislosti genů.

Klíčová slova: Procedury mnohonásobného testování, microarray, genové exprese

Title: Multidimensional statistics and applications to study genes

Author: Mgr. Peter Bubelíny

Department: Department of probability and mathematical statistics

Supervisor: prof. Lev Klebanov, DrSc., KPMS MFF UK

Abstract: Microarray data of gene expressions consist of thousands of genes and just some tens of observations. Moreover, genes are highly correlated between themselves and contain systematic errors. Hence the magnitude of these data does not afford us to estimate their correlation structure. In many statistical problems with microarray data, we have to test some thousands of hypotheses simultaneously. Due to dependence between genes, *p*-values of these hypotheses are dependent as well. In this work, we compared convenient multiple testing procedures reasonable for dependent hypotheses. The common manner to make microarray data more uncorrelated and partially eliminate systematic errors is normalizing them. We proposed some new normalizations and studied how different normalizations influence hypotheses testing. Moreover, we compared tests for finding differentially expressed genes or gene sets and identified some interesting properties of some tests such as bias of two-sample Kolmogorov-Smirnov test and interesting behavior of Hotelling's test for dependent components of observations. In the end of this work, we proposed test for testing independence of genes.

Keywords: Multiple testing procedures, microarray, gene expressions

# Contents

# Preface

DNA microarrays are part of promising class of biotechnologies that allow the measure of expression levels of thousands of genes simultaneously. These expressions can be utilized for example to find changes between different biological statements. An important question in microarray experiment is the identification of genes which are associated with a response (e.g. dose of a drug, time, treatment/control) and covariates (e.g. survival time, clinical outcome) of interest. This leads to the problem of multiple hypotheses testing. That is a testing of null hypothesis for each gene simultaneously.

Another approach to the microarray data, but not considered in this work, can be for example clustering. The primary goal of clustering is grouping genes with similar expression patterns (e.g. separating cancerous from noncancerous genes). Similar expression patterns can offer insights into various biological processes (*D'haeseleer et al.* (2000)).

Probably the first paper using microarray experiment was presented by Shena in his paper *Schena et al.* (1995). Since then, there has been a growing number of publication and some successes about this topic. For example in 1999, there was shown that patients with leukemia can be accurately classified into two known subgroups with using just gene expressions (*Golub et al.* (1999)). Another success turn up in 2001 (*Sorlie* (2001)) when researchers identified five patterns of gene expression levels in breast cancer and showed that they correspond to different types of disease with different prognosis. In paper *Zembutsu et al.* (2002), they use microarrays with more than 23,000 features to predict the response to anti-cancer drugs in terms of efficacy and toxicity on a group of patients. In 2003, there were identified 158 genes associated with pancreatic cancer that were differentially expressed with comparison to people with a healthy pancreas. Another success was reached in *Petty et al.* (2006). There was discovered a gene that has highly different expression between cancer patients who respond to the chemotherapy treatment and patients who did not respond to the chemotherapy treatment.

### Problems

This work concerns the use of microarrays in a comparative experiment which is desired to compare gene expressions between two datasets. In this issue, one would like to identify which of several thousands of candidate genes have had their expression levels changed, that is identify which genes are differentially expressed.

There are at least two main problems that make the work with gene expression data complicated. The first one is the number of genes. We often have several thousands of genes. Consider that we test all hypotheses at significance level $\alpha = 0.05$. Then we reject about 5% of true hypotheses from several thousands. In result, we determine several hundreds of non-differentially expressed genes as differentially expressed. Detailed investigation of genes costs a lot of time and money. Therefore, we cannot afford so many type I error (false positive) hypotheses. For the cost of power, this problem can be partially solved by using proper multiple testing procedure (chapter 3) and/or grouping some genes together (chapter 6).

The second problem is that gene expressions are highly correlated between genes. *Klebanov and Yakovlev* (2007) studied various microarray data sets. They found out that the average of correlation coefficients between genes ranged from 0.84 to 0.97. Because we usually have only a few tens of observations, we cannot estimate the covariance structure of gene expression data. Hopefully, there exists some normalizations, which make gene expression data almost uncorrelated and they can help us to handle this problem (chapter 5).

### State of arts

This work consists of 8 separate chapters. Each chapter deals with different problems of microarray experiment or connected problems. Therefore, the results of each chapter do not directly depend on the results of the other chapters. Hence, every chapter can be read just with basic knowledge of this problem and without knowledge of the other chapters.

Chapter 1 serves as a brief introduction to genetics, microarray experiments and describes the process of obtaining and preprocessing of microarray data. There have been a lot of papers dealing with these problems in detail, e.q. *Yakovlev et al.* (2013) or *Göhlmann and Talloen* (2009). Because this chapter just summarize some known things, it is based on citations and there is no our own contributions.

Chapter 2 concerns the choice of test for finding differentially expressed genes between two states of observations (e.g. healthy/ill, two kinds of some disease). Usually, the $t$-test or tests based on $t$-statistic are used (see *Dudoit et al.* (2003)). In paper *Zinger et al.* (1989), there was considered another test, called $N$-test, for this problem. Moreover, two-sample Kolmogorov-Smirnov test is reasonable test for this problem as well.

In *Gordon and Klebanov* (2010), they proved that for $n = m$ there exists $\alpha \in (0, 1)$ such that two-sample Kolmogorov-Smirnov test is unbiased at level $\alpha$ against two-sided alternative $F \neq G$. In this work, we extended this theorem for one-sided alternatives $A_1 : F \leq G$ or $A_2 : F \geq G$. We discovered that for each $n \neq m$ there exists $\alpha \in (0, 1)$ such that this test is unbiased against one-sided alternative. When we considered two-sided alternative again, we discovered that for $n \neq m$ there exists $\alpha \in (0, 1)$ and the distribution $F_\alpha$ such that this test is biased against this alternative distribution $F_\alpha$. Furthermore, we discovered that this test need not to be unbiased against this alternative for another choice of $\alpha^* \neq \alpha$. These results for two-sample Kolmogorov-Smirnov test were published in *Bubeliny* (2013a). At the end of chapter 3, we compared the power of $t$-test, $N$-test and two-sample Kolmogorov-Smirnov test and found out that $N$-test serves as a good alternative for $t$-test in case of violence of normality.

Chapter 3 deals with the properties of $p$-values of tests about gene expressions. It is well known that gene expressions are highly correlated between genes, see *Klebanov and Yakovlev* (2007). Therefore, $p$-values of tests about gene expressions are dependent as well. However, we have seen no papers concerning the behavior of $p$-values. Therefore, we showed how histogram of such $p$-values can look like and how can be changed their structure in case of using some normalization such a proportion of gene expressions. Results of this chapter were published in *Bubeliny* (2008).

While working with gene expression data we often need to test a lot of hypotheses simultaneously. If we used classical approach with significance level $\alpha = 5\%$ we would expect to reject about 5% of all true hypotheses. It means too much hypotheses with type I error. In case of some thousands of genes this number is unacceptable. Therefore, chapter 4 deals with multiple testing procedures that eliminate this problem. The most known procedure is Bonferroni procedure *Bonferroni* (1936). This procedure controls FWER (family wise error rate), that is the probability of commit at least one type I error. Moreover, this procedure controls expected number of type I errors at predefined level $\alpha$. But this procedure is generally considered as to be too conservative. Therefore, different approach was proposed by *Benjamini and Hochberg* (1995). They derived procedure that controls FDR (false discovery rate) - expected proportion of type I errors among rejected hypotheses. Another approach, called empirical Bayes approach, closely related to FDR was proposed in *Efron* (2003). An overview of different multiple testing procedures can be found for example in *Dudoit and van der Laan* (2008). In our work, we performed an extensive simulation study to compare some of these procedures. Our study showed that it is the principle of multiple testing procedures and not the principle of Bonferroni procedure that makes this procedure to be considered too conservative. Therefore, Bonferroni procedure should not be underestimated.

It was mentioned before, that gene expressions are highly correlated between genes. Moreover, there are many sources of systematic variations in microarray experiments that affect measure of gene expressions. Therefore, microarray data are normalized

*Yang et al.* (2002). But do these normalizations help in deciding process of finding differentially expressed genes? In chapter 5, we performed simulation study to find how some normalizations affect testing null hypotheses. We discovered that common used procedures like quantile normalization and global normalizations result in finding too many false positives (hypotheses with type I error). Therefore, we proposed normalization based on $\delta$-sequence (see *Klebanov and Yakovlev* (2007)). We showed that our normalization finds reasonable number of false positives and detects more true positives (invalid null hypotheses) than we would test without normalizing. Partial results of this chapter were published as *Bubeliny* (2013b).

Gene expression data consists of some thousands of genes and therefore we are expected to test some thousands of hypotheses simultaneously. To decrease this number, we can group some genes into gene sets and test the equality of distributions of these sets (e.g. *Barry et al.* (2008)). In that case, we are dealing with two-sample multidimensional problem. The most popular tests for this problem are Hotellings test, N-test and tests derived from marginal $t$-statistics. Our pre-study of Hotelling's test for genes sets showed strange behavior of this test. Therefore, in chapter 6 we looked at this test from theoretical point of view and found some interesting results. We discovered that this test does not need to reach the best power in case that all marginal distributions are shifted. In case of strong dependence of components of sample vectors, better power is achieved in case of one marginal shift between samples than in case of all marginal distributions are equally shifted. Moreover, for highly dependent components the best power is achieved when about half of marginal distributions are equally shifted. These results about Hotelling's test were published in *Bubeliny* (2011). At the end of this chapter, we compared the power of this test with $N$-test and two tests based on marginal $t$-statistics. These results confirmed different behavior of Hotelling's test with comparison with another considered tests.

Type of dependence between genes can be very helpful. For example, consider that there exists a gene which influences group of genes such that if this gene is differentially expressed the whole group of genes will be differentially expressed. Then we could investigate just this gene in detail instead of the group of genes. In *Klebanov et al.* (2006), they defined a type of dependence between genes called type A dependence. In *Lim et al.* (2010), they defined another type of dependence between genes called hidden regulator dependence (HRD). They numerically demonstrated that HRD is easily mistaken for type A dependence. In this work, we defined test for testing independence between genes. By using this test, we discovered that there exist a lot of pairs of genes with type A dependence. On the other hand, our results showed that there is just small proportion of these pairs of genes with hidden regulator dependence. Therefore, type A dependence is more frequent among pairs of genes than HRD.

Finally, our results are summarized in chapter 8.

# Chapter 1

# Introduction

Statistician working with gene expression data should know how these data were gained and how these data were preprocessed before they get to his/her hand. Because the beginning of these data roots in deoxyribonucleic acid (DNA) that contains the organisms complete hereditary information, we start this chapter with biological introduction into human genome. Thereafter, we describe how these data are preprocessed until they get into the statistician hand. Furthermore, we briefly describe basic approaches of analyzing microarray data and how these data can be represented. More detailed introduction to the microarray data analysis can be found for example in *Yakovlev et al.* (2013) or *Göhlmann and Talloen* (2009). At the end of this chapter, we describe HYPERDIP and TEL data for childhood leukemia that we use through this work.

## 1.1 Human Genome

A eukaryote is an organism (e.g. human being, animals, plants . . .) whose cells contain a nucleus and other structures enclosed within membranes. Most eukaryotic organisms have billions of individual cells. Almost all of these cells contain the entire genome for that organism. This genome carries complete hereditary information in the form of deoxyribonucleic acid (DNA).

The human genome consists of 23 pairs of chromosomes. Each chromosome is made of chains of DNA. DNA consists of molecules that are wrapped around each other in a structure known as a double helix. Genes are essentially segments of the DNA structure. In other words a gene is a section of DNA. In humans, there are about 27,000 of genes. The information contained in the gene is transcribed into a messenger ribonucleic acid (mRNA). Then this mRNA molecule leaves the nucleus of the cell and it is transcribed into a protein (translation process). This process is known as gene expression.

## 1.2 DNA Microarrays

DNA microarrays (also commonly known as DNA chip or biochip) are small solid supports (for example microscope slides, silicon chips). The idea behind microarrays is measuring the amount of the different types of mRNA molecules in a cell and thus indirectly measure the expression levels of the genes. Each DNA microarray spot contains a specific DNA sequences, known as probes (also reporters or oligos). These probes are complementary to the specific mRNA molecules that correspond to the specific targeting genes. These mRNA molecules, which have been previously labeled with fluorescent dye, should hybridize with those probes. The amount of hybridization is then measured by the amount of fluorescence. It is usually done by scanner and the results are subsequently analyzed by computer. A spot with brighter fluorescence means that the gene represented by this spot has higher expression level.

Two kind of DNA microarrays are used nowadays: oligonucleotide arrays and cDNA arrays.

The most commonly used DNA microarray is oligonucleotide array called GeneChip manufactured by Affymetrix (http://www.affymetrix.com). Each array contains hundreds of thousands of probe spots and each of these spots contains millions of copies of an individual 25 base long DNA oligonucleotide.

In cDNA microarrays, each spot corresponds entirely to a specific gene. The probes are generally hundreds of bases long and measure complementary DNA (cDNA). The expression level is then given by the measure of how much cDNA hybridize to its corresponding spot. Moreover, two separate samples are hybridized to the same array at the one time. One of these samples is a control sample and the second is a sample of interest (e.g. cancer tissue) and they are labeled with different dye. Expression level of a given gene is then measured by the difference in intensity level. Scanner which reads cDNA microarrays produces a TIFF image. These images are processed by image analysis software.

## 1.3 Data preprocessing

The goal of preprocessing microarray data is to remove undesired sources of variations. The raw microarray data of different probes are noisy. In *Holloway et al.* (2006), there was shown that suitable preprocessing step is crucial to obtain reliable data.

The whole preprocessing procedure begins when probe intensities are stored in the image of scanned microarrays where each pixel of image can have some discrete level of gray. After image acquisition, each probe is identified by the grip placed on the top of scanned image and is represented by a set of pixels. From this set of pixels the overall probe intensity is calculated. A typical first preprocessing step is background correction that aims to remove non-biological contributions to the intensities such as background patterns across arrays, unspecific binding of the transcript, etc. In Affymetrix software, the square microarray image is divided into 16 squares. For each particular square, the background intensity is declared to be the second percentile of all probe intensities in this square. The value that is subtracted from a given probeset is weighted average of 16 background intensities where weights depend on the distance of probeset from considered squares.

Then usually follows the base 2 logarithmic transformation of probeset intensities. This is done due to the fact that $\log_2$-transformation makes the microarray intensity distribution more symmetric (see e.g. *Chen et al.* (2007)). The second reason is that the intensity variations usually increase with intensities. Furthermore, a biological side effect of $\log_2$-transformation is that this transformation converts multiplicative effects into additive effects.

The following preprocessing step is normalization that makes different samples of an experiment comparable among themselves. The main goal of normalization is to remove systematic differences between chips. If an experiment is done perfectly, there is no need of normalization (there are no systematic variations). Different normalizations involve various assumptions on data properties. Therefore, we should use them carefully. The most known normalizations are quantile normalization and global normalization. The wider summary of different normalizations and their comparison can be found for example in *Yang et al.* (2002).

Microarray technology can measure the whole genome at once. But not all genes are expected to be expressed. Furthermore, there exist some genes that cannot or have to be differentially expressed (for maintenance cell). Therefore, to reduce dimensionality, some genes can be omitted.

## 1.4 Data analysis

Gene expression data are useful only if one can extract meaningful information from them. Appropriate analysis can discover unknown properties of genes. On the other hand, inferior analysis can lead to wrong results and mislead the researchers. Depending on the goal of the analysis, various statistical methods should be chosen. There are numerous technics that concern about microarray data analysis.

### 1.4.1 Classification

There are two important approaches of classification of microarray data. The first one is the discrimination between different known cell patterns, e.g. between tumor and normal tissue. The second one is the identification of unknown cell types or conditions, e.g. new subclass of existing subclass of tumors. In the statistical literature, they are known as discrimination (supervised) and clustering (unsupervised) methods. Clustering methods are more appropriate if cell classes are not known in advance. On the other hand, discriminant methods are preferred if the classes are known.

### 1.4.2 Discrimination methods

Suppose, that we have $n$ multivariate (consist of some genes) samples $X_1, \ldots, X_n$ of gene expression data. Suppose that there exist K classes of cell profiles. The goal of discriminant analysis is to define $K$ disjoint subset $A_k$, $k = 1, \ldots, K$ of sample space such that for $X_i \in A_k$ the predicted class is $k$. These subsets are built from observations which are known to belong to one of the considered class. The most known discrimination methods are for example Fisher linear discriminant analysis, maximum likelihood discriminant analysis, nearest neighbor and classification trees. Detailed description of these methods can be found for example in *Dudoit et al.* (2002), where these methods were moreover compared.

### Clustering methods

Clustering method is a technique by which genes or samples are grouped based on pairwise similarities between genes/samples. For these methods, there are two important choices to be done. The first one is the distance measure (similarity) between two elements. The most known distance measures are for example Euclidian distance, Manhattan distance and Pearson correlation coefficient. The second one (called linkage) defines how similar elements need to be in order to be assigned into the same cluster. The examples of linkages are for example nearest neighbor, furthest neighbor and average linkage. Clustering methods are divided into two categories: hierarchical methods and partitioning methods. Hierarchical methods build successive cluster using previous clusters. Partitioning methods are based on minimization of heterogeneity of clusters. Detailed description of clustering methods can be found for example in *Quackenbush* (2001) or *Göhlmann and Talloen* (2009).

### 1.4.3 Finding differentially expressed genes

The common goal of microarray study is to identify differentially expressed genes under specific conditions. It is done by testing equality of distribution of gene expressions. Detection of differentially expressed genes depends on the design of experiment, on the choice of testing statistics and the predefined significant level. There can be two experimental conditions or many. If we consider two experimental condition we dealing with two-sample problem. Reasonable tests for such conditions can be for example *t*-test. If we consider more than two experimental conditions then the reasonable test can be for example analysis of variance. Detail description of tests to identify differentially expressed genes can be found for example in *Cui and Churchill* (2003). The choice of significant level depends on the type I error rate (e.q. family wise error rate, false discovery rate etc.) and on the multiple testing procedure (e.q Bonferroni procedure, Benjamini-Yekutieli procedure etc.). Detailed summary of multiple testing procedures and various type I error rates can be found for example in *Dudoit and van der Laan* (2008).

## 1.5 Microarray data

Microarray experiment, which produces gene expressions of $m$ distinct genes, can be represented by random vector $\mathbf{X} = (x_1, \ldots, x_m)'$ with mutually dependent components. Consider that we have $n$ samples (slides) of $\mathbf{X}$. Therefore, we can represent microarray data for $m$ genes from $n$ slides by $m$ x $n$ matrix $X = \{\mathbf{X_1}| \ldots |\mathbf{X_n}\} = \{x_{i,j}\}_{i,j=1}^{m,n}$, where $x_{i,j}$ is the gene expression level for $i$-th gene from $j$-th slide.

## 1.6 HYPERDIP and TEL data

During this work, we use HYPERDIP and TEL data for childhood leukemia. We do not try to study these data. We just use them to verify our simulated results on real data. These data are free to obtain from St. Jude Chilren's research hospital (http://www.stjude.org). They were observed on children's patient of this hospital. These data were proceed by Affymetrix microarray. Both datasets consist of non-normalized data and have 7084 genes. For HYPERDIP data, there were obtained 88 slides and for TEL data 79 slides. More details about processing of these data can be found in supplementary information of *Yoeh et al.* (2002) where these data were analyzed. The data, which we use, can be found in the supplement of this work.

In the following chapters, we would like to use *t*-test or Hotelling's test for these data. It was previously mentioned that $\log_2$ transformation of gene expression data is expected to have approximately normal distribution. Therefore, we should verify whether $\log_2$ transformation of HYPERDIP and TEL data have the normal distribution. To do this, we just verify whether each gene expression level has normal distribution. We test normality according to one-sample Kolmogorov-Smirnov test and Shapiro-Wilk test. Histograms of *p*-values of these tests for both HYPERDIP and TEL data are in figure

1.1. Because we are testing 7084 hypotheses we cannot use significance level $\alpha = 5\%$. In case of using it we are expected to reject about 0.05 x 7084 = 354.2 true hypotheses. Instead of it, according to Bonferonni procedure (see Section 4.2 for more details), we should use significance level $\alpha^* = 0.05/7084$. Number of rejected hypotheses according to significance level $\alpha$ and $\alpha^*$ are in table 1.1. Although we reject some hypotheses by Shapiro-Wilk test (one for HYPERDIP data and five for TEL data, respectively), their proportion among all genes is too small. Therefore, we can say, that $\log_2$ transformation of these data are approximately normal distributed.

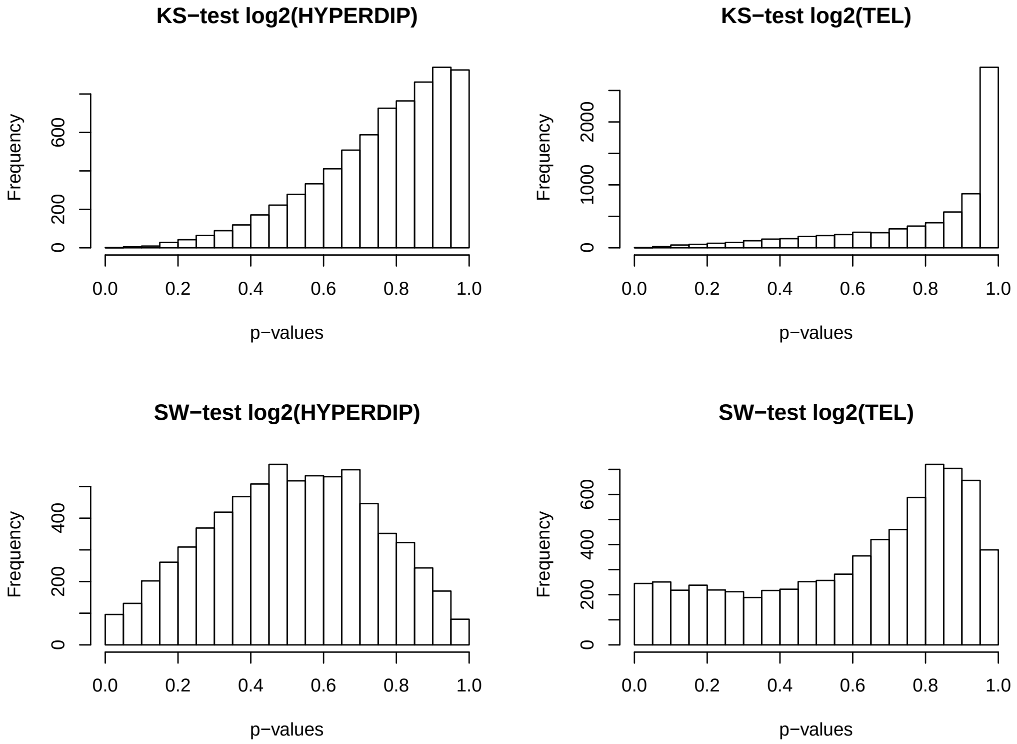

Figure 1.1: $p$-values of one-sample Kolmogorov-Smirnov test and Shapiro-Wilk test for normality of genes from HYPERDIP and TEL data.

| # of rejected hypotheses | HYPERDIP | | TEL | |
|---|---|---|---|---|
| | $\alpha$ | $\alpha^*$ | $\alpha$ | $\alpha^*$ |
| KS-test | 1 | 0 | 3 | 0 |
| SW-test | 96 | 1 | 245 | 5 |

Table 1.1: Number of rejected hypotheses of normality by one-sample Kolmogorov-Smirnov test and Shapiro-Wilk test at significance level $\alpha = 0.05$ and $\alpha^* = 0.05/7084$ for HYPERDIP and TEL data.

# Chapter 2

# Tests for gene expression data

By working with microarray data, we often need to test the hypothesis of equality of mean value of two samples or equality of distributions of two samples. The common used test for such problem is $t$-test. An alternative to this test (especially if we are interested in equality of distributions or there is a violation of normality) can be for example $N$-test or two-sample Kolmogorov-Smirnov test. At the beginning of this chapter, we describe $N$-test that was derived in *Zinger et al.* (1989). In what follows, we discuss the biasedness of one-sample Kolmogorov-Smirnov test that can be used e.g. to verify some assumptions. Moreover, we show some interesting properties about biasedness of two-sample Kolmogorov-Smirnov test that were published in *Bubeliny* (2013a). At the end of this chapter, we compare the power of $t$-test, $N$-test and two-sample Kolmogorov-Smirnov test.

## 2.1 $N$-test

Let $\mu$ and $\nu$ be two probability measures defined on the Euclidean space $R^d$. For testing the hypothesis $H : \mu = \nu$, *Zinger et al.* (1989) derive the distribution free test, called $N$-test.

Let $L(x, y)$ be a strictly negative definite kernel, this is $\sum_{i,j=1}^{s} L(x_i, x_j)h_i h_j \leq 0$ for any $x_1, \ldots, x_s$ and $h_1, \ldots, h_s$, $\sum_{i=1}^{s} h_i = 0$ with equality if and only if all $h_i = 0$. Define

$$\begin{aligned} N(\mu, \nu) &= 2\int_{R^d}\int_{R^d} L(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) \\ &\quad - \int_{R^d}\int_{R^d} L(x, y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) - \int_{R^d}\int_{R^d} L(x, y)\mathrm{d}\nu(x)\mathrm{d}\nu(y), \end{aligned}$$

then $\sqrt{N(\mu, \nu)}$ is a metric in the space of all probability measures on $R^d$.

Suppose, that $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_{n_1})'$ and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{n_2})'$ are two independent $d$-

dimensional vector samples, consisting of $n_1$ and $n_2$ observations, from $\mu$ and $\nu$, respectively. Then, the empirical counterpart of $N(\mu,\nu)$ is given by

$$\hat{N}(\mathbf{x},\mathbf{y}) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} 2L(\mathbf{x}_i,\mathbf{y}_j) - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} L(\mathbf{x}_i,\mathbf{x}_j) - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} L(\mathbf{y}_i,\mathbf{y}_j).$$

As a strictly negative definite kernel $L$ we will use Euclidian distance that is defined by $L(\mathbf{x}_i,\mathbf{y}_j) = \sqrt{\sum_{l=1}^{d}(x_{il}-y_{jl})^2}$ for each $i = 1,\ldots,n_1$ and $j = 1,\ldots,n_2$. For this kernel, $N(\mu,\nu) = 0$ if and only if $\mu = \nu$. The higher $\hat{N}(\mathbf{x},\mathbf{y})$ the stronger evidence to reject hypothesis $H$: $\mu = \nu$.

Exact distribution of statistic $N(.,.)$ is not known, therefore we will estimate the $p$-value of $N$-test by permutations according to the following algorithm.

**Algorithm 2.1.**

1. *Compute* $\hat{N}(\mathbf{x},\mathbf{y})$.
2. *Let* $\mathbf{z}$ *be a pooled sample of d-dimensional samples* $\mathbf{x}$ *and* $\mathbf{y}$. *In other words,* $\mathbf{z}$ *can be rewritten as* $\mathbf{z} = (\mathbf{z}_1,\ldots,\mathbf{z}_{n_1+n_2})' = (\mathbf{x}_1,\ldots,\mathbf{x}_{n_1},\mathbf{y}_1,\ldots,\mathbf{y}_{n_2})' = (\mathbf{x},\mathbf{y})'$.
3. *Permutate d-dimensional vectors, the components of* $\mathbf{z}$, *to gain new sample* $\mathbf{z}^{(i)} = (\mathbf{z}_1^{(i)},\mathbf{z}_2^{(i)})'$ *(with sample sizes of* $n_1$ *and* $n_2$*) and compute* $\hat{N}(\mathbf{z}_1^{(i)},\mathbf{z}_2^{(i)})$.
4. *Repeat step* 3 *K-times.*
5. *Estimated p-value of N-test is given by*

$$\hat{p} = \frac{1}{K}\sum_{i=1}^{K} I_{[\hat{N}(\mathbf{x},\mathbf{y})\leq\hat{N}(\mathbf{z}_1^{(\mathbf{i})},\mathbf{z}_2^{(\mathbf{i})})]},$$

*where* $I_{[.]}$ *is the indicator function.*

The disadvantage of $N$-test is that it is too time consumable. Due to the large number of genes, we cannot perform too many permutations. It limits the choice of $K$. For gene expression data it can be set just to some thousands. Therefore, the $p$-values estimated by this test can be inaccurate (especially for low $p$-values).

The simplest and the most common case is when $x_1,\ldots,x_{n_1}$ and $y_1,\ldots,y_{n_2}$ are two independent samples having one-dimensional distributions with the distribution functions $F$ and $G$, respectively. We would like to test the hypothesis $H : F = G$ against the alternative $A : F \neq G$. Then $L(x,y) = |x-y|$ and $N$-test statistic for this hypothesis is given by

$$\hat{N}(\mathbf{x},\mathbf{y}) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} |x_i - y_j| - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} |x_i - x_j| - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} |y_i - y_j|.$$

To find out if this test hold the significance level $\alpha = 0.05$ we performed simple simulation. We simulated data from normal, log-normal and uniform distribution and from 5000 repetitions we estimated the significance level of this test for different number of observations as proportion of rejections among all repetitions. Estimates of $p$-values of $N$-test were based on 1000 permutations. In table 2.1, there are estimates of level $\alpha$ in case of equal number of observations ($n_1 = n_2 = 10, 15, 20, 30, 50, 75, 100$) in both samples. In table 2.2, there are results of simulation for nonequal number of observations ($n_1 = 10, 15, 20, 30, 50, 75, 100$ and $n_2 = 12, 25, 60, 95$). We can see that in all simulated cases, the estimates of $\alpha$ are near 0.05. Therefore, we can say that $N$-test hold significance level $\alpha$.

| $n_1 = n_2$ | 10 | 15 | 20 | 30 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|
| normal | 0.049 | 0.053 | 0.049 | 0.047 | 0.050 | 0.054 | 0.049 |
| log-normal | 0.053 | 0.050 | 0.053 | 0.049 | 0.052 | 0.057 | 0.054 |
| uniform | 0.055 | 0.055 | 0.046 | 0.050 | 0.050 | 0.050 | 0.047 |

Table 2.1: Estimate of significance level $\alpha = 5\%$ of $N$-test for normal, log-normal and uniform distributed samples with sample sizes $n_1 = n_2 = 10, 15, 20, 30, 50, 75, 100$.

## 2.2 One-sample and two-sample Kolmogorov-Smirnov test

One-sample and two-sample Kolmogorov-Smirnov test are distance-based tests. Consider that we have a sample $x_1, \ldots, x_n$ from the distribution with unknown distribution function $F$. Let $\hat{F}_n(x)$ denotes its empirical distribution function. Based on this sample, one would like to test the hypothesis $H_1 : F = F_0$ against the alternative $A_1 : F \neq F_0$, where $F_0$ is fixed distribution function. Then one-sample Kolmogorov-Smirnov test is based on statistic

$$D_n = d(\hat{F}_n(x), F_0(x)) = \sup_x |\hat{F}_n(x) - F_0(x)|.$$

The hypothesis $H_1$ is rejected at level $\alpha$ if and only if $D_n$ is greater than critical value $\delta_{\alpha,n}$ of this test.

Consider now that we have another sample (independent with the first) $y_1, \ldots, y_m$ from the distribution having unknown distribution function $G$. We would like to test the hypothesis $H_2 : F = G$ against the alternative $A_2 : F \neq G$. Then two-sample Kolmogorov-Smirnov test is based on statistic

$$D_{n,m} = d(\hat{F}_n(x), \hat{G}_m(x)) = \sup_x |\hat{F}_n(x) - \hat{G}_m(x)|.$$

| $n_1$ | 10 | 15 | 20 | 30 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|
| $n_2$ | | | | normal | | | |
| 12 | 0.046 | 0.053 | 0.051 | 0.046 | 0.045 | 0.049 | 0.052 |
| 25 | 0.051 | 0.050 | 0.043 | 0.052 | 0.049 | 0.050 | 0.046 |
| 60 | 0.045 | 0.049 | 0.052 | 0.050 | 0.047 | 0.047 | 0.049 |
| 95 | 0.051 | 0.053 | 0.053 | 0.055 | 0.049 | 0.050 | 0.046 |
| $n_2$ | | | | log-normal | | | |
| 12 | 0.049 | 0.043 | 0.045 | 0.051 | 0.044 | 0.052 | 0.054 |
| 25 | 0.049 | 0.048 | 0.052 | 0.048 | 0.052 | 0.050 | 0.050 |
| 60 | 0.051 | 0.056 | 0.050 | 0.054 | 0.053 | 0.055 | 0.051 |
| 95 | 0.053 | 0.050 | 0.053 | 0.053 | 0.047 | 0.047 | 0.047 |
| $n_2$ | | | | uniform | | | |
| 12 | 0.052 | 0.051 | 0.051 | 0.053 | 0.051 | 0.049 | 0.055 |
| 25 | 0.052 | 0.049 | 0.057 | 0.051 | 0.052 | 0.051 | 0.051 |
| 60 | 0.058 | 0.053 | 0.048 | 0.048 | 0.047 | 0.050 | 0.052 |
| 95 | 0.053 | 0.046 | 0.058 | 0.054 | 0.047 | 0.049 | 0.054 |

Table 2.2: Estimate of significance level $\alpha = 5\%$ of $N$-test for normal, log-normal and uniform distributed samples with sample sizes $n_1 = 10, 15, 20, 30, 50, 75, 100$ and $n_2 = 12, 25, 60, 95$.

The hypothesis $H_2$ is rejected at level $\alpha$ if and only if $D_{n,m}$ is greater than its predefined critical value $\delta_{\alpha,n,m}$.

Computation of $p$-values of both versions of Kolmogorov-Smirnov test can be found for example in *Hajek et al.* (1999).

It should be kept in mind that both Kolmogorov-Smirnov tests do not depend on any monotonic transformation of samples. In other words, if we transform both samples (by the same monotonic transformation) to samples with the distribution functions $F'$ and $G'$, respectively then

$$\sup_x |\hat{F}_n(x) - F_0(x)| = \sup_x |\hat{F}'_n(x) - F'_0(x)|$$

and

$$\sup_x |\hat{F}_n(x) - \hat{G}_m(x)| = \sup_x |\hat{F}'_n(x) - \hat{G}'_m(x)|,$$

where $F'_0$ is the transformed distribution function of $F_0$.

### 2.2.1 Biasedness of one-sample Kolmogorov-Smirnov test

Although one-sample tests have only few direct applications for gene expression data, they are often used to verify various assumptions. Therefore, we consider one-sample

Kolmogorov-Smirnov test here (specially its biasedness).

Recall that a test is said to be unbiased at level $\alpha$ if

1. it has significance level $\alpha$
2. for all alternative distributions the power of this test is greater or equal to $\alpha$.

The test is said to be unbiased if it is unbiased at all levels $\alpha \in (0, 1)$. Finally, the test is said to be biased if it is not unbiased. Specially, the test is biased at level $\alpha$ against alternative $G$ if it is an level $\alpha$ test and $P(\text{reject } H|G) < \alpha$. The distribution $G$ is said to be the most biased distribution of test for hypothesis $H$ at significance level $\alpha$ if $G$ minimizes the probability of rejection hypothesis $H$ at level $\alpha$ among all distributions, that is $P_\alpha(\text{reject } H|G) \leq P_\alpha(\text{reject } H|G') \;\; \forall G'$.

Each continuous distribution can be transformed to uniform (0,1) distribution. Because one-sample Kolmogorov-Smirnov test does not depend on any monotonic transformation we can assume, without loss of generality, that we have independent sample $X = (x_1, \ldots, x_n)'$ from distribution with continuous distribution function $F$ with $\operatorname{supp} F \subseteq [0, 1]$. Now we would like to test the hypothesis $H : F = F_0$ against the alternative $A : F \neq F_0$, where $F_0$ is the the $(0, 1)$ uniform distribution (for simplicity we will write just uniform distribution) given by

$$F_0(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases} . \tag{2.1}$$

Now let $\delta$ and $\beta$ be such constants that $0 < \delta \leq \delta_{\alpha,n}$ and $\beta > 1$ and define $\delta^*$ by $\delta^* = \delta\frac{\beta-1}{\beta}$. In paper *Massey* (1950), Massey found out that one-sample Kolmogorov-Smirnov test is biased against two sided alternative with distribution function defined by

$$G(x) = \begin{cases} 0 & \text{if } x < \delta^* \\ \beta x - \delta(\beta - 1) & \text{if } \delta^* \leq x < \delta \\ x & \text{if } \delta \leq x < 1 - \delta \\ \beta x - (\beta - 1)(1 - \delta) & \text{if } 1 - \delta \leq x < 1 - \delta^* \\ 1 & \text{if } x \geq 1 - \delta^* \end{cases} . \tag{2.2}$$

Confidence set for empirical distribution function of one-sample Kolmogorov-Smirnov test is given by a closed ball $B(F_0; \delta_{\alpha,n})$ of radius $\delta_{\alpha,n} > 0$ centered at $F_0$ in the metric of all distribution functions with the Kolmogorov distance. In figure 2.1, there are plotted distribution functions of $F_0$ (left one) and $G$ (right one) for $\delta = 0.2$ and $\beta = 2$ together with confidence sets with $\delta_{\alpha,n} = 0.2$ for one-sample Kolmogorov-Smirnov test. The biasedness of Kolmogorov-Smirnov test against alternative distribution $G$ is now evident from the following theorem of *Gordon and Klebanov* (2010).

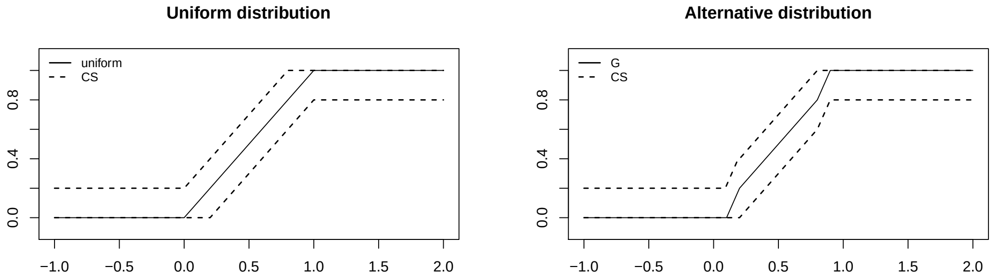

Figure 2.1: Distribution functions of uniform and alternative distribution (solid line) and their confidence intervals (dashed line) of one-sample Kolmogorov-Smirnov test with $\delta_{\alpha,n} = 0.2$.

**Theorem 2.2.1.** *Suppose that for some* $0 < \alpha < 1$ *there exists a continuous distribution function* $F_a$ *such that*

$$B(F_a; \delta_{\alpha,n}) \subset B(F_0; \delta_{\alpha,n}) \tag{2.3}$$

*and for difference of sets* $B(F_0; \delta_{\alpha,n})$ *and* $B(F_a; \delta_{\alpha,n})$ *holds*

$$P_{F_a}(\hat{F}_{na} \in B(F_0; \delta_{\alpha,n})/B(F_a; \delta_{\alpha,n})) > 0, \tag{2.4}$$

*where* $\hat{F}_{na}$ *is empirical distribution function of* $F_a$*. Then one-sample Kolmogorov-Smirnov test is biased against the alternative* $F_a$*.*

*Proof.* Let $x_1, \ldots, x_n$ be independent identically distributed variables from the distribution $F_a$ with empirical distribution function $\hat{F}_{na}$. Then

$$P_{F_a}(\hat{F}_{na} \in B(F_a; \delta_{\alpha,n})) \geq 1 - \alpha.$$

From (2.3) and (2.4) we have

$$P_{F_a}(\hat{F}_{na} \in B(F_0; \delta_{\alpha,n})) > 1 - \alpha.$$

That is

$$P(\text{reject } H|A \text{ is true}) = P_{F_a}(d(\hat{F}_{na}; F_0) > \delta_{\alpha,n}) < \alpha.$$

□

From this theorem, it is obvious that distribution functions given by (2.2) are not the only ones that make this test biased. For testing equality with uniform distribution, one-sample Kolmogorov-Smirnov test is biased at level $\alpha$ to all distributions with continuous distributions functions (say $F_a$) lower than $F_0$ for $x \leq \delta_{\alpha,n}$ and greater than $F_0$ for $x \geq 1 - \delta_{\alpha,n}$ with restriction to $P_{F_a}(\hat{F}_{na} \in B(F; \delta_\alpha)/B(F_a; \delta_\alpha)) > 0$.

### 2.2.2 Simulations of alternative distributions

In this chapter, we will need to simulate data from an alternative distribution $G$. It can be directly done from sample $X$ having uniform distribution by $Y = G^{-1}(X)$, where $G^{-1}$ is the inversion function of $G$. More precisely, for the distribution function given by (2.2), this is

$$Y(x) = G^{-1}(x) = \begin{cases} \frac{x+(\beta-1)\delta}{\beta} & \text{if } x < \delta \\ x & \text{if } \delta \leq x < 1-\delta \\ \frac{x+(\beta-1)(1-\delta)}{\beta} & \text{if } 1-\delta \geq x \end{cases} . \tag{2.5}$$

To find out how strong the biasedness of one-sample Kolmogorov-Smirnov test is we performed simple simulation. We set the number of observations $n$ to be $n = 50, 100$ and parameter $\beta$ to be $\beta = 2, 5, 10, 50, 1000$. For $n = 50$ we set $\delta$ to be $0.1, 0.15, 0.187$ and for $n = 100$ to be $0.08, 0.11, 0.135$. Each simulation was repeated 100000 times. For each setting we simulated random samples $X = (x_1, \ldots, x_n)'$ from the uniform distribution and from these samples (for better comparison) we calculated samples $Y = (y_1, \ldots, y_n)'$ according to (2.5). For each sample $X$ and $Y$, we performed one-sample Kolmogorov-Smirnov test and estimated how many times it rejected the hypothesis $H$ at level $\alpha = 5\%$. In table 2.3, there are showed the differences between estimates of level $\alpha$ for sample $X$ and the estimates of power for sample $Y$. For example, for $n = 50$, $\delta = 0.187$ and $\beta = 2$ the estimate of level $\alpha$ is equal to 0.05039, the estimate of power for sample $Y$ is equal to 0.04962. Therefore, the difference is equal to 0.00077. All differences of estimates in our simulation are nonnegative and they are larger with increasing $\delta$. It confirms that the one-sample Kolmogorov-Smirnov test is not unbiased.

If sample $X$ is transformed to sample $Y$ according to (2.5), it does not necessarily mean that $p$-value of one-sample Kolmogorov-Smirnov test is changed. Figure 2.2 illustrates $p$-values for 2000 simulation (top row) and there is 1063 (53.15%) from these $p$-values that were changed between samples $X$ and $Y$. We can see that $p$-values for

| | $n = 50$ | | | $n = 100$ | | |
|---|---|---|---|---|---|---|
| | $\delta = 0.1$ | $\delta = 0.15$ | $\delta = 0.187$ | $\delta = 0.08$ | $\delta = 0.11$ | $\delta = 0.135$ |
| $\beta = 2$ | 0.00004 | 0.00024 | 0.00077 | 0.00000 | 0.00002 | 0.00005 |
| $\beta = 5$ | 0.00004 | 0.00028 | 0.00073 | 0.00001 | 0.00004 | 0.00009 |
| $\beta = 10$ | 0.00002 | 0.00033 | 0.00087 | 0.00000 | 0.00002 | 0.00015 |
| $\beta = 50$ | 0.00003 | 0.00022 | 0.00100 | 0.00000 | 0.00003 | 0.00011 |
| $\beta = 1000$ | 0.00005 | 0.00035 | 0.00097 | 0.00000 | 0.00004 | 0.00017 |

Table 2.3: Table of difference between estimates of level $\alpha$ for sample $X$ having uniform distribution and estimate of power for sample $Y$ having distribution function (2.2) by one-sample Kolmogorov-Smirnov test at level $\alpha = 0.05$.

samples from alternative distribution are upper bounded by some constant (in this case about 0.75) and the majority of changed $p$-values are greater than another constant (in this case about 0.5). The upper bound of $p$-values for alternative samples is due to empirical distribution function of alternative distribution that is ever equal to zero in $\delta^*$, because the minimum of $Y$ is never less than $\delta^*$. That is $\hat{G}_n(\delta^*) = 0$. Therefore, $D_n = \sup_x |\hat{G}_n(x) - F_0(x)| \geq \delta^*$.

### 2.2.3 Modifications of one-sample Kolmogorov-Smirnov test

Consider, that there is a real threat that the testing distribution is the distribution given by (2.2). Now we know that one-sample Kolmogorov-Smirnov test can lead to wrong decisions. However, what can we do? In the rest of this subsection, we propose three modifications of this test that could help. In these modifications we use one-sample Kolmogorov-Smirnov test together with another tests. According to Union-intersection principle developed by *Roy* (1957) we can write the main hypothesis $H$ as an intersection of partial hypotheses $H_\tau$, where $\tau$ is some set of hypotheses, that is $H = \bigcap_\tau H_\tau$. On the other hand, we reject hypothesis $H$ if we reject at least one partial hypothesis. On the grounds of holding significance level of hypothesis $H$ and according to Bonferroni inequality, for partial hypotheses $H_\tau$ we use significance level $\alpha^* = \alpha/|\tau|$, where $|\tau|$ is number of partial hypotheses.

### First modification

We know that the minimum of sample from the distribution $G$ is never lower than $\delta^*$. Now consider, that we have sample $X = (x_1, \ldots, x_n)$ from the uniform distribution. The minimum of sample $X$ is greater than or equal to $\delta^*$ with probability given by

$$P(\min X \geq \delta^*) = P(\forall x_i \geq \delta^*) = \prod_{i=1}^{n} P(x_i \geq \delta^*) = (1 - \delta^*)^n,$$

which for a reasonable $n$ and $\delta^*$ is too small. Moreover, the same idea can be applied to the maximum of the sample. It leads us to the first modification of one-sample Kolmogorov-Smirnov test.

In this modification, we use the minimum and the maximum of the sample $X$. Therefore, we need to know the critical values of these statistics. Let $\delta_{\text{min}}$ and $\delta_{\text{max}}$ be the critical values for these statistics at level $\alpha^*$. For the minimum of $X$ we have

$$P(\min X < \delta_{min}) = 1 - P(\min X \geq \delta_{min}) = 1 - (1 - \delta_{min})^n = \alpha^*,$$

therefore $\delta_{min} = 1 - (1 - \alpha^*)^{1/n}$. The computation for maximum is analogous and we have $\delta_{max}^n = (1 - \alpha^*)^{1/n}$. Because we are going to use three tests in one test, we set $\alpha^* = \alpha/3$. The test of modification one is summarized by the following algorithm.

**Algorithm 2.2.**

1. *Compute p-value (denoted by $p_{ks}$) of one-sample Kolmogorov-Smirnov test for the sample $X$.*
2. *Compute the minimum and the maximum of the sample $X$.*
3. *The hypothesis H is rejected if and only if:*

$$(\min X < 1 - (1-\alpha^*)^{1/n}) \text{ or } (\max X > (1-\alpha^*)^{1/n}) \text{ or } (p_{ks} < \alpha^*)$$

.

## Second modification

Distribution functions of the uniform and the alternative distribution $G$ are different from 0 to $\delta$ and from $1-\delta$ to 1. Therefore, we can use one-sample Kolmogorov-Smirnov test for small values and large values of $X$ separately. Moreover, we omit the values from the "middle" of $X$.

If $\delta$ in distribution $G$ given by (2.2) is less than or equal to critical value of one-sample Kolmogorov-Smirnov test $\delta_{\alpha,n}$ then the one-sample Kolmogorov-Smirnov test is biased against this alternative. Hence, as a sample $X_1$ we take normalized values of $X$ which are lower than $\delta_{\alpha,n}$, that is

$$X_1 = \{x_i/\delta_{\alpha,n}; x_i < \delta_{\alpha,n}, i = 1, \dots, n\}. \tag{2.6}$$

Alternatively, we create a sample of "large" values of X from normalized values of $X$ which are larger than $1-\delta_{\alpha,n}$, that is

$$X_2 = \{(1-x_i)/\delta_{\alpha,n}; x_i > 1-\delta_{\alpha,n}, i = 1, \dots, n\}. \tag{2.7}$$

Such defined samples $X_1$ and $X_2$ are independent and if the hypothesis is true then both of these samples have the uniform distribution. Therefore, on each of these samples we can use one-sample Kolmogorov-Smirnov test separately. It leads us to the second modification of one-sample Kolmogorov-Smirnov test, which is given by the following algorithm.

**Algorithm 2.3.**

1. *Compute p-value (denoted by $p_{ks}$) of one-sample Kolmogorov-Smirnov test for the sample $X$.*
2. *Compute p-values (denoted by $p_{min}$ and $p_{max}$) of one-sample Kolmogorov-Smirnov test for the samples $X_1$ and $X_2$ given by (2.6) and (2.7), respectively.*
3. *The hypothesis H is rejected if and only if at least of one of three considered tests is rejected at level $\alpha/3$, that is $p_{ks} < \alpha/3$ or $p_{min} < \alpha/3$ or $p_{max} < \alpha/3$.*

Note that sample $X_1$ or $X_2$ can be empty with positive probability and therefore one-sample Kolmogorov-Smirnov test cannot be computed. If the hypothesis $H$ is true, then this probability is equal to $(1 - \delta_{\alpha,n})^n$ which is too small (i.e for n=50 and $\alpha$=0.05 it is equal to approximately 2.3x$10^{-5}$). Therefore, we set $p$-value of empty sample to be zero.

## Third modification

The third modification is similar to the second modification. The only difference is that in this modification, samples $X_1$ and $X_2$ are joined together and there is created just one sample $X_3$. If the hypothesis $H$ is true, $X_1$ and $X_2$ are independent and they both have uniform distribution and therefore $X_3$ as well. Hence, we can apply one-sample Kolmogorov-Smirnov test on sample $X_3$. The third modification of one-sample Kolmogorov-Smirnov test is summarized by the following algorithm.

**Algorithm 2.4.**

1. *Compute p-value (denoted by $p_{ks}$) of one-sample Kolmogorov-Smirnov test for the sample $X$.*
2. *Compute p-values (denoted by $p_{mix}$) for the sample $X_3$.*
3. *The hypothesis $H$ is rejected if and only if at least one of two considered tests is rejected at level $\alpha/2$, that is $p_{ks} < \alpha/2$ or $p_{mix} < \alpha/2$.*

Again, if the sample $X_3$ is empty, we set $p_{mix} = 0$.

### 2.2.4 Power of modified one-sample Kolmogorov-Smirnov tests

Previously, we proposed three modifications of one-sample Kolmogorov-Smirnov that should improve this test against the alternative $G$. In this section, we performed simulation study, in which we verified if these tests hold nominal level $\alpha = 0.05$ of the main hypothesis $H$ and we compared their power.

In order to verify if our three tests hold nominal level $\alpha$, we simulated random samples from uniform distribution. We set the number of observations $n$ to be 50,100 and 1000. We performed 10000 repetitions and as the estimate of level $\alpha$ we took the proportion of rejected hypotheses between these 10000 simulations. From table 2.4, we can see that all three modifications hold nominal level $\alpha$.

The power of our three modifications is still questionable. Therefore, we performed simulation to compare the power of these modifications. We considered number of observations $n = 50$ and 100, we set parameter $\beta = 2, 5, 10, 50$. The last parameter we did need to set was $\delta$. For each setting of $n$ and $\beta$ we set fifty different equidistant values of

|  | K-S test | Mod 1 | Mod 2 | Mod 3 |
|---|---|---|---|---|
| $n = 50$ | 0.0494 | 0.0473 | 0.0490 | 0.0482 |
| $n = 100$ | 0.0486 | 0.0449 | 0.0479 | 0.0511 |
| $n = 1000$ | 0.0475 | 0.0479 | 0.0450 | 0.0480 |

Table 2.4: Estimate of level $\alpha = 0.05$ of one-sample Kolmogorov-Smirnov test and its three modifications.

$\delta$. Each simulation was repeated 5000 times.

Figure 2.3 shows the results of our simulation. It can be seen, that each of our modifications improves one-sample Kolmogorov-Smirnov test against the alternative $G$. For $n = 50$ and each $\beta$ the best power has the modification three. For $n = 100$ and $\beta = 2$, the first modification has the highest power, for larger $\beta$ the third modification has similar power to the first one.

Let consider another alternative distribution. We assume that alternative distribution function increases polynomial between 0 and $\delta$ and between $1-\delta$ and 1. For simplicity, for $x$ from 0 to $\delta$ we will consider $m$-degree polynomial $P_1$ with $a_0 = \ldots = a_{m-1} = 0$ and $a_m = \beta_1$, where constant $\beta_1$ is such that polynomial $P_1$ satisfies $P_1(0) = 0$ and $P_1(\delta) = \delta$. For $x$ from $1 - \delta$ to 1 we consider $m$-degree polynomial $P_2$ with $a_1 = \ldots = a_{m-1} = 0$ and $a_0 = \gamma$ and $a_m = \beta_2$, where constants $\gamma$ and $\beta_2$ are such that $P_2(1 - \delta) = 1 - \delta$ and $P_2(1) = 1$. It leads to the alternative distribution function given by

$$G_m(x) = \begin{cases} 0 & \text{if } x < 0 \\ \beta_1 x^m & \text{if } 0 \leq x < \delta \\ x & \text{if } \delta \leq x < 1-\delta \\ \beta_2 x^m + \gamma & \text{if } 1-\delta \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}. \tag{2.8}$$

At first, we performed simulation study in order to confirm that one-sample Kolmogorov-Smirnov test is biased against the alternative $G_m$. We set $n = 50, 100$, $m = 2, 5, 10, 50, 1000$, three different values of $\delta$ and we performed 100000 simulation. In table 2.5, there are differences between estimates of level $\alpha = 5\%$ for sample having uniform distribution and the estimate of power for sample having distribution function $G_m$ created from these uniformly distributed samples. Each of these differences is nonnegative. That acknowledges that one-sample Kolmogorov-Smirnov test is biased against the alternative $G_m$.

Now we compare the power of our modifications for the alternative $G_m$. We considered number of observations $n = 50$ and 100, we set the degree of polynomials $m = 2, 5, 10, 50$. For parameter $\delta$ we took fifty different equidistant values of $\delta$. Each simulation was repeated 5000 times.

| | $n = 50$ | | | $n = 100$ | | |
|---|---|---|---|---|---|---|
| | $\delta = 0.1$ | $\delta = 0.15$ | $\delta = 0.187$ | $\delta = 0.08$ | $\delta = 0.11$ | $\delta = 0.135$ |
| $m = 2$ | 0.00009 | 0.00029 | 0.00071 | 0.00000 | 0.00003 | 0.00005 |
| $m = 5$ | 0.00003 | 0.00037 | 0.00088 | 0.00001 | 0.00005 | 0.00017 |
| $m = 10$ | 0.00007 | 0.00023 | 0.00082 | 0.00000 | 0.00004 | 0.00017 |
| $m = 50$ | 0.00002 | 0.00029 | 0.00074 | 0.00001 | 0.00002 | 0.00015 |
| $m = 1000$ | 0.00001 | 0.00034 | 0.00093 | 0.00000 | 0.00004 | 0.00015 |

Table 2.5: Table of difference between estimates of level $\alpha$ for sample having the uniform distribution and estimates of power for samples having the distribution given by (2.8) by one-sample Kolmogorov-Smirnov test at level $\alpha = 0.05$.

The results of this simulation for modified tests are in figure 2.4. For $n = 50$ the third modification has the highest power. For $n = 100$ it is hard to say if modification one or modification three is the best. It depends on $m$ and $\delta$ because lines of power are crossed for these two modifications.

### 2.2.5 Some notes on biasedness of two-sample Kolmogorov-Smirnov test

Consider, that $x_1, \dots, x_n$ and $y_1, \dots, y_m$ are two independent identically distributed samples having distributions with continuous distribution functions $F$ and $G$, respectively. We would like to test the hypothesis $H : F = G$ against the alternative $A : F \neq G$. Then two-sample Kolmogorov-Smirnov test is based on statistic

$$D_{n,m} = \sup_x |\hat{F}_n(x) - \hat{G}_m(x)|,$$

where $\hat{F}_n(x)$ and $\hat{G}_m(x)$ are the empirical distribution functions of $F$ and $G$. The hypothesis $H$ is rejected for large values of $D_{n,m}$.

At first, we should realize that statistic $D_{n,m}$ of two-sample Kolmogorov-Smirnov test has discrete distribution. Therefore, $p$-values for this test have a discrete distribution as well. For example, consider the case $n = m = 50$. Then the test statistic $D_{n,m}$ can take just 50 different values $1/n, 2/n, \dots, 1$. For statistic $D_{n,m} = 0.26$ the $p$-value is equal to 0.0678 and for the next value $D_{n,m} = 0.28$ the $p$-value is equal to 0.0392. Testing at level $\alpha = 0.05$ could be a bit confusing because the power of this test is constant for each value $\alpha \in [0.0392, 0.0678)$. There exists a distribution $G$ such that power of two-sample Kolmogorov-Smirnov test at level $\alpha = 0.05$ is equal to 0.045. Such a distribution does not meet requirements of definition of unbiasedness for $\alpha = 0.05$ though the power of this test is higher than exact level of this test equal to 0.0392. To precise the idea of unbiasedness for tests with discrete test statistic we will consider just discrete values of significance level $\alpha$.

It should be kept in mind that the two-sample Kolmogorov-Smirnov test does not depend on any monotonic transformation of samples. If we transform both samples (by the same monotonic transformation) to samples with distribution functions $F'$ and $G'$, respectively then $\sup_x |\hat{F}_n(x) - \hat{G}_m(x)| = \sup_x |\hat{F}'_n(x) - \hat{G}'_m(x)|$. Therefore, without loss of generality, we assume that $F$ is the uniform distribution given by

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases} . \tag{2.9}$$

In *Gordon and Klebanov* (2010), there was proved that for $n = m$ there exists $\alpha \in (0, 1)$ such that two-sample Kolmogorov-Smirnov test is unbiased at level $\alpha$ against two-sided alternative $F \neq G$. If we consider just one-sided alternatives $A_1 : F \leq G$ or $A_2 : F \geq G$ we can extend this finding to $n \neq m$.

**Theorem 2.2.2.** *Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ be independent samples from distribution $F$ and $G$. Then for arbitrary $n, m \in N$, there exists $\alpha \in (0, 1)$ such that two-sample Kolmogorov-Smirnov test of hypothesis $H : F = G$ against one-sided alternative $A_1 : F \leq G$ or $A_2 : F \geq G$ is unbiased at level $\alpha$.*

*Proof.* Without loss of generality, we assume that the first sample $x_1, \ldots, x_n$ is from the uniform distribution.
Firstly, we consider only the alternative $A_1 : F \leq G$. For this alternative, the Kolmogorov-Smirnov statistic is given by

$$D^*_{n,m} = \sup_{x \in (0,1)} (\hat{F}_n(x) - \hat{G}_m(x)),$$

where $\hat{F}_n$ and $\hat{G}_m$ are the empirical distribution functions of $F$ and $G$. The hypothesis $H$ is rejected for small values of $D^*_{n,m}$. Consider $\alpha$ such small, that we reject the hypothesis $H$ for $D_{n,m}$ equals to minus one. It occurs if and only if the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ satisfy

$$\max(y_1, \ldots, y_m) < \min(x_1, \ldots, x_n). \tag{2.10}$$

The probability of this event is given by

$$n \int_0^1 (1 - x)^{n-1} G^m(x) dx. \tag{2.11}$$

Moreover, $G(x)$ must be monotone and $G(x) \geq x$ because we consider alternative $A_1 : F \leq G$. Therefore, the function $(1 - x)^{n-1} G^m(x)$ of integral (2.11) attains its minimum for $G(x) = x$. This integral represents probability of rejection of hypothesis at level

$\alpha$ if alternative $G$ is true and it is minimized for $F = x = G(x)$. Hence, two-sample Kolmogorov-Smirnov test is unbiased at level $\alpha$.

The proof for the alternative $A_2 : F \geq G$ is similar. We take $\alpha$ such small, that we reject hypothesis if and only if $D_{n,m} = 1$. The inequality (2.10) change to

$$\max(x_1, \ldots, x_n) < \min(y_1, \ldots, y_m)$$

and probability of this event is then given by

$$n \int_0^1 x^{n-1}(1 - G(x))^m dx \tag{2.12}$$

For alternative $A_2$ we have $G(x) \leq x$. Hence, integral (2.12) is minimized for $G(x) = x$. It proves the theorem. □

The result of this theorem does not mean that two-sample Kolmogorov-Smirnov test is unbiased against one-sided alternative. It only says that there exists small level $\alpha$ for which this test is unbiased. In the following theorem we show that for $n \neq m$ two-sided Kolmogorov-Smirnov test is not unbiased against two-sided alternative.

**Theorem 2.2.3.** *Let $x_1, \ldots, x_n$ be i.i.d from uniform distribution with distribution function $F$ and $y_1, \ldots, y_m$ be i.i.d. from distribution having distribution function $G$. If $n \neq m$ then there exists $\alpha \in (0, 1)$ such that two-sample Kolmogorov-Smirnov test of hypothesis $H : F = G$ is biased against alternative with the distribution function*

$$G(x) = \frac{(\frac{x}{1-x})^{\frac{n-1}{m-1}}}{1 + (\frac{x}{1-x})^{\frac{n-1}{m-1}}}. \tag{2.13}$$

*Proof.* Consider $\alpha$ such small, that we reject the hypothesis $H$ if and only if

$$D_{n,m} = \sup_x |\hat{F}_n(x) - \hat{G}_m(x)| = 1.$$

That is, the samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ have to satisfy

$$\max(y_1, \ldots, y_m) < \min(x_1, \ldots, x_n) \text{ or } \max(x_1, \ldots, x_n) < \min(y_1, \ldots, y_m). \tag{2.14}$$

The probability of this event is given by

$$n \int_0^1 \left((1 - x)^{n-1} G^m(x) + x^{n-1}(1 - G(x))^m\right) dx.$$

Substitute $G(x)$ by $y$ and let the derivative (according to $y$) of function

$$(1 - x)^{n-1} y^m + x^{n-1}(1 - y)^m,$$

be equal to zero. It leads us to the equation

$$(\frac{y}{1-y})^{m-1} = (\frac{x}{1-y})^{n-1}.$$

Therefore, the probability of event (2.14) is not minimized for $F(x) = G(x) = x$ but for

$$G(x) = \frac{(\frac{x}{1-x})^{\frac{n-1}{m-1}}}{1 + (\frac{x}{1-x})^{\frac{n-1}{m-1}}}.$$

□

Some examples of distribution functions given by (2.13) are in figure 2.5. Although we found out that two-sample Kolmogorov-Smirnov test is biased against alternative (2.13) we showed it just for very small $\alpha$. Let denote this smallest level $\alpha$ by $\alpha_1$. Then $\alpha_1$ can be directly computed by

$$\alpha_1 = n\int_0^1 ((1-x)^{n-1}x^m + x^{n-1}(1-x)^m)\,dx = 2nm\frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m+1)}, \tag{2.15}$$

where $\Gamma(.)$ denotes Euler gamma function. For example if $n = 10$ and $m = 11$ then $\alpha_1$ is equal to 5.67x$10^{-6}$.

All previous results are considered for Kolmogorov-Smirnov statistic $D_{n,m} = 1$. Let us consider the second highest value of this statistic. For $n > m$ it is equal to $1 - 1/n$ and for $n < m$ it is equal to $1 - 1/m$, respectively. We denote by $\alpha_2$ the significance level $\alpha$ such that we reject two-sample Kolmogorov-Smirnov test if and only if $D_{n,m} \geq \max(1 - 1/n, 1 - 1/m)$.

Firstly, assume that $n > m \geq 2$ and consider that $D_{n,m} = 1 - 1/n$. This can occur if and only if these samples are such that

$$x_{(1)} < \ldots < x_{(n-1)} < y_{(1)} < x_{(n)}$$

or

$$x_{(1)} < y_{(m)} < x_{(2)}, \ldots < x_{(n)}.$$

Together with the case when $D_{n,m} = 1$, that is $x_{(n)} < y_{(1)}$ or $y_{(m)} < x_{(1)}$, we have that $D_{n,m}$ is greater or equal to $1 - 1/n$ if and only if $x_{(n-1)} < y_{(1)}$ or $y_{(m)} < x_{(2)}$. It leads us to the probability of rejecting the hypothesis at level $\alpha_2$

$$\begin{aligned} P(D_{n,m} \geq 1 - 1/n) &= P(\forall_j y_j > x_{(n-1)}) + P(\forall_j y_j < x_{(2)}) \\ &= n(n-1)\int_0^1 (x^{n-2}(1-x)(1-G(x))^m \\ &\quad + x(1-x)^{n-2}G^m(x))\,dx. \end{aligned} \tag{2.16}$$

As in the proof of the previous theorem let $G(x) = y$ and let the derivative (according to $y$) of the integrand of (2.16) equal to zero. It leads to solve the equation

$$(\frac{y}{1-y})^{m-1} = (\frac{x}{1-x})^{n-3}.$$

The solution $y$ as a function of $x$ is given by

$$y = G(x) = \frac{(\frac{x}{1-x})^{\frac{n-3}{m-1}}}{1 + (\frac{x}{1-x})^{\frac{n-3}{m-1}}}. \tag{2.17}$$

Now assume that $2 \le n < m$ and consider $D_{n,m} = 1 - 1/m$. This can be true if and only if

$$y_{(1)} < \ldots < y_{(m-1)} < x_{(1)} < y_{(m)}$$

or

$$y_{(1)} < x_{(n)} < y_{(2)}, \ldots < y_{(m)}.$$

Therefore, the probability of event $D_{n,m} \ge 1 - 1/m$ is equal to

$$\begin{aligned} P(D_{n,m} \ge 1 - 1/m) &= P(D_{n,m} = 1 - 1/m) + P(D_{n,m} = 1) \\ &= nm \int_0^1 ((1-x)^{n-1} G^{m-1}(x)(1 - G(x)) \\ &\quad + x^{n-1}(1 - G(x))^{m-1} G(x))\, dx \\ &\quad + n \int_0^1 ((1-x)^{n-1} G^m(x) + x^{n-1}(1 - G(x))^m)\, dx. \end{aligned} \tag{2.18}$$

As before let $G(x) = y$ and let the derivative of the integrand of (2.18) according to $y$ be equal to zero, leading to the equation

$$(\frac{y}{1-y})^{m-3} = (\frac{x}{1-x})^{n-1}.$$

Therefore, the distribution function of the most biased distribution of two-sample Kolmogorov-Smirnov test at level $\alpha_2$ is given by

$$y = G(x) = \frac{(\frac{x}{1-x})^{\frac{n-1}{m-3}}}{1 + (\frac{x}{1-x})^{\frac{n-1}{m-3}}}. \tag{2.19}$$

**Remark 2.2.4.** *If $n = 3$ and $m = 2$ or $n = 2$ and $m = 3$ then the most biased distribution of two-sample Kolmogorov-Smirnov test is discrete distribution given by probabilities $P(y = 0) = P(y = 1) = \frac{1}{2}$ or $P(y = \frac{1}{2}) = 1$, respectively.*

Consider $G(x) = x$ then level $\alpha_{n,m} = \alpha_2$ is given (according to (2.16) and (2.18)) by

$$\alpha_2 = 2nmk\frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m+1)} = k\alpha_1, \tag{2.20}$$

where $k = \min(n+1, m+1)$. The distribution functions (2.17) and (2.19) are $S$-shaped (see figure 2.5). Although these distribution functions are not identical and not equal to (2.13), some interesting results can be found. If $|n - m| = 2$ then (2.17) and (2.19) change to $G(x) = x$. It means that the distribution which minimizes (2.16) and (2.18) is uniform distribution. It leads us to the following theorem.

**Theorem 2.2.5.** *Let $\alpha_{n,m}$ be given by (2.20). If $n = m + 2$ or $n = m - 2$ then two-sample Kolmogorov-Smirnov test is unbiased at level $\alpha_{n,m}$. However, if $n \neq m$ and $|n - m| \neq 2$ then Kolmogorov-Smirnov test is biased at level $\alpha_{n,m}$.*

*Proof.* Because of $\alpha_{n,m} = \alpha_2$, the distribution functions of the most biased distribution of this test at level $\alpha_2$ are given by (2.17) and (2.19). For $|n - m| = 2$ they change to $G(x) = x = F(x)$. It means that the uniform distribution minimize the probability of rejection hypothesis $F = G$ against alternative $F \neq G$ at level $\alpha_2$ if and only if $|n - m| = 2$. □

**Remark 2.2.6.** *If $|n - m| = 1$ then two-sample Kolmogorov-Smirnov test is not biased against the distribution functions (2.17) and (2.19) at level $\alpha_1$.*

Let denote by $\mathscr{A}_\alpha$ the set of distributions for which two-sample Kolmogorov-Smirnov test is biased at level $\alpha$, it is

$$\mathscr{A}_\alpha = \{G : P(\text{reject } H \text{ at level } \alpha | \text{alternative } G \text{ is true}) < \alpha\}.$$

For different levels $0 < \alpha < \alpha^*$, one would expect that there is some subset relation between $\mathscr{A}_\alpha$ and $\mathscr{A}_{\alpha^*}$. However, it is not generally true. According to the theorem 2.2.5 there exist $G_\alpha$ such that $G_\alpha \in \mathscr{A}_\alpha$ and $G_\alpha \notin \mathscr{A}_{\alpha^*}$. On the other hand, from remark 2.2.6 we have that there exists $G^*_\alpha$ such that $G^*_\alpha \notin \mathscr{A}_\alpha$ and $G^*_\alpha \in \mathscr{A}_{\alpha^*}$. Therefore, in general $\mathscr{A}_\alpha$ is not subset of $\mathscr{A}_{\alpha^*}$ and vice versa.

Previous result can be quite simply generalized to $\alpha_3$ (the third smallest $\alpha$) in case of $n > 2m$ or $2n < m$. Adding the probability of the even $D_{n,m} = 1 - 2/m$ or $D_{n,m} = 1 - 2/n$ to the (2.16) or (2.18) leads us to the most biased distributions at level $\alpha_3$ given by

$$G_3(x) = \frac{(\frac{x}{1-x})^{\frac{n-5}{m-1}}}{1 + (\frac{x}{1-x})^{\frac{n-5}{m-1}}} \qquad \text{if } n > 2m \tag{2.21}$$

or

$$G_3(x) = \frac{(\frac{x}{1-x})^{\frac{n-1}{m-5}}}{1 + (\frac{x}{1-x})^{\frac{n-1}{m-5}}} \qquad \text{if } m > 2n. \tag{2.22}$$

In this case, $\alpha_3$ is given by

$$\alpha_3 = 2k_2nm\frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m+1)} = k_2\alpha_1,$$

where

$$k_2 = \frac{\min((m+2)(m+1),(n+2)(n+1))}{2}.$$

If $n = m + 4$ or $m = n + 4$ then $G_3(x) = x$. Together with condition $n > 2m$ or $m > 2n$ we have that for $n = 6, m = 2$ or $n = 2, m = 6$ the two-sample Kolmogorov-Smirnov test is unbiased at level $\alpha_3 = 3/7$ and for $n = 7, m = 3$ or $n = 3, m = 7$ the two-sample Kolmogorov-Smirnov test is unbiased at level $\alpha_3 = 1/6$.

Computing the power of two-sample Kolmogorov-Smirnov test for another relation of $n$ and $m$ at level $\alpha_3$ is not such simple due to the fact that it has to be solved by double integration. Therefore, in such cases, finding the most biased distribution is much more complicated and it is not considered here.

The $\alpha$'s considered so far are too small in case we have some tens of observations in each sample. Therefore, we performed the following simulation to look if two-sample Kolmogorov-Smirnov test is biased against the distribution (2.13) at level $\alpha \approx 0.05$. We set the number of observations $n$ for the first sample be $n = 10, 20, 50, 100$ and the number of observations $m$ for the second sample be $m = 11, 15, 21, 51, 101$. As a distribution of the first sample, we consider uniform distribution and for the second sample, we consider two distributions: the uniform distribution and distribution given by (2.13). We performed 10000 repetitions and computed the difference between the estimate of power if the second sample is from alternative distribution and the estimated level $\alpha$ if the second sample is from uniform distribution. The results of this simulation are in table 2.6. We can see that for all considered $n$ and $m$ the estimate of difference is greater than 0. It means that two-sample Kolmogorov-Smirnov test is not biased against alternative (2.13) at level $\alpha \approx 0.05$ for the chosen parameters $n$ and $m$.

| $\alpha = 5\%$ | m=11 | m=15 | m=21 | m=51 | m=101 |
|---|---|---|---|---|---|
| $n = 10$ | 0.0034 | 0.0144 | 0.0320 | 0.4153 | 0.7290 |
| $n = 20$ | 0.0291 | 0.0087 | 0.0016 | 0.2784 | 0.9170 |
| $n = 50$ | 0.4071 | 0.3403 | 0.2715 | 0.0001 | 0.5291 |
| $n = 100$ | 0.9070 | 0.9189 | 0.9190 | 0.4557 | 0.0001 |

Table 2.6: Difference between estimate of power for alternative $G$ given by (2.13) and estimate of level $\alpha$ of two-sample Kolmogorov-Smirnov test.

## 2.3 Comparison of power of $N$-test, $t$-test and two-sample Kolmogorov-Smirnov test

One of the goals of microarray experiment is to find differently expressed genes between two different groups of patients or two stages of some disease. This is closely related with two-sample test of hypothesis of equality of distributions of gene expression levels between these two groups. Therefore, $N$-test and two-sample Kolmogorov-Smirnov test seem to be useful in such situation. On the other hand, many biologists are interested in change of mean value of gene expression instead of difference in distribution. Moreover, $\log_2$ expressions are considered to have approximately normal distribution, therefore, $t$-test seems to be useful as well.

One can be interested in which test is the best. Therefore, we performed simulation to compare the power of $N$-test, $t$-test and two-sample Kolmogorov-Smirnov test for some specific alternatives. We set the number of observations in each sample to be $n_1 = n_2 = 10, 20, 50, 100$. We compared samples from $N(0, 1)$ and $N(\mu, 1)$; $\log N(0, 1)$ and $\log N(\mu, 1)$; $U(0, 1)$ and $U(0, 1) + \mu$; $N(0, 1)$ and $N(0, \sigma^2)$. In each comparison we considered 10 equidistant values $\mu$ and $\sigma$ which differ from case to case.

Results of these simulations are in figure 2.6. We can see that $N$-test has better power than two-sample Kolmogorov-Smirnov test in all simulated cases. If there is a change only in mean value then $t$-test performs slightly better than $N$-test. But if there is a change in variance (log-normal and normal with change in variance cases) then $N$-test has far better power than $t$-test. Therefore, $N$-test seems to be better than $t$-test for samples with different covariance structure. On the other hand, $N$-test is much more time-consuming than $t$-test. This fact should be likewise considered in choosing the test we are going to use.

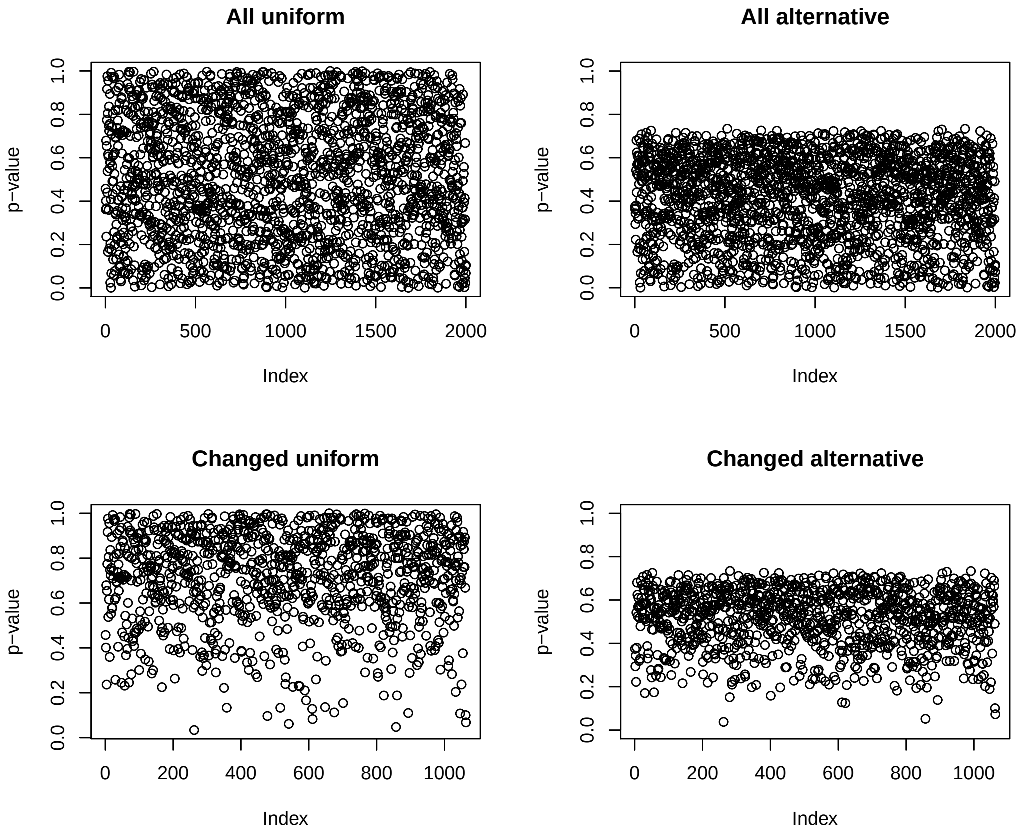

Figure 2.2: $p$-values of one-sample Kolmogorov-Smirnov test of sample-size $n = 50$ for uniform and alternative distribution with $\delta = 0.187$ and $\beta = 2$. In the top row, there are all $p$-values for 2000 simulations. In the bottom row, there are 1063 $p$-values that changed between samples $X$ and $Y$ for these 2000 simulations.

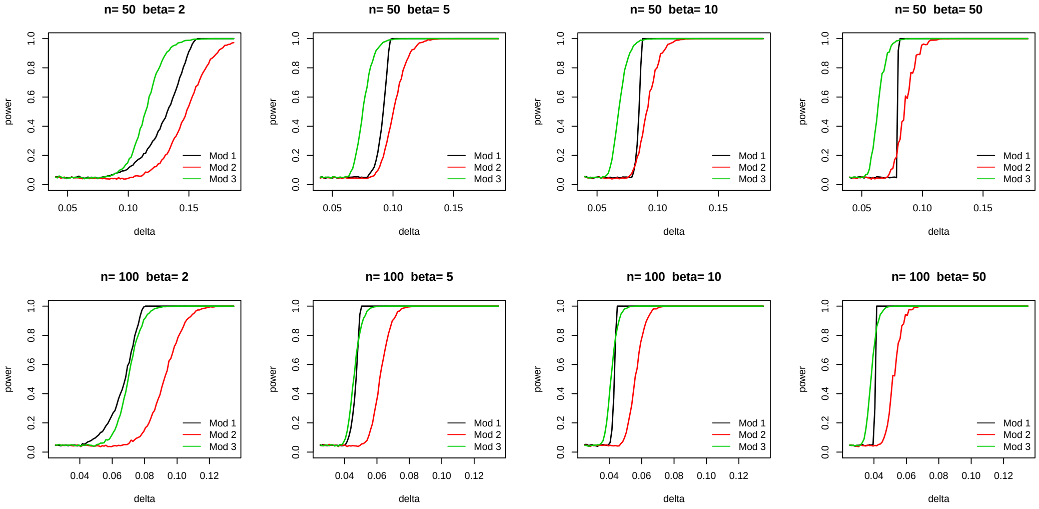

Figure 2.3: Estimate of power of our three modifications for sample having distribution function $G$ given by (2.2) with sample size $n = 50, 100$.

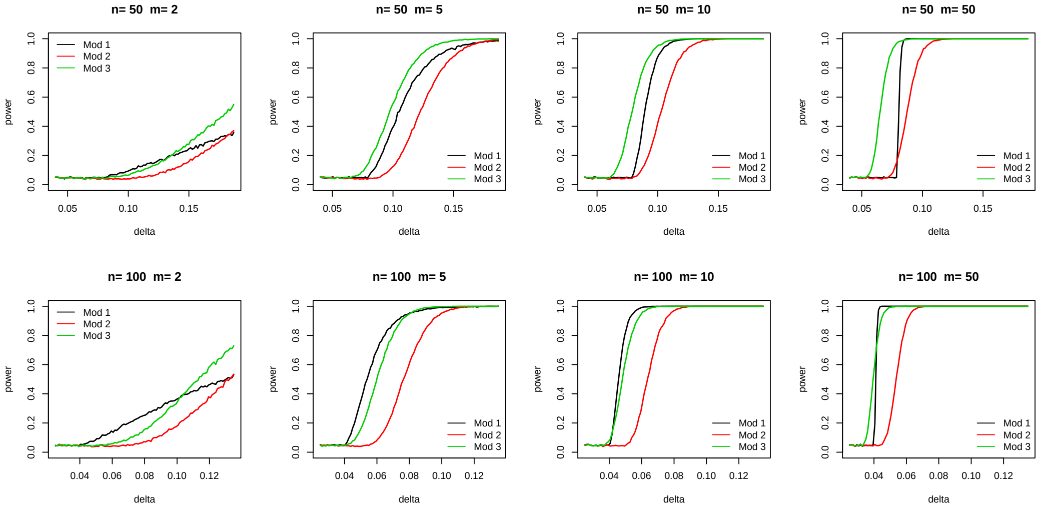

Figure 2.4: Estimate of power of our three modifications for sample having distribution function $G_m$ given by (2.8) with sample size $n = 50, 100$ .

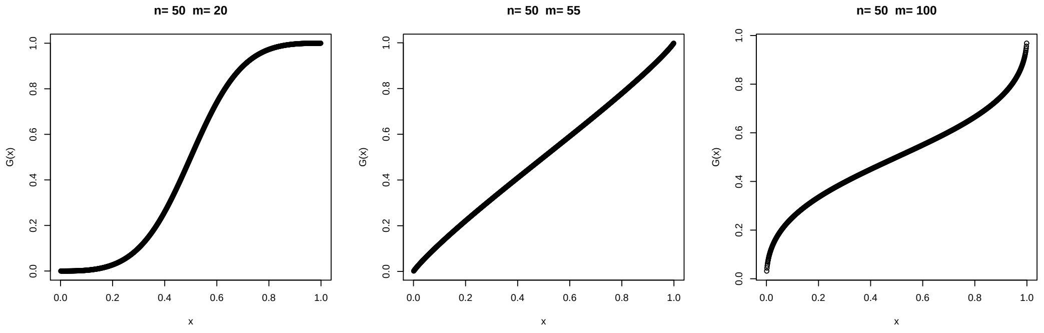

Figure 2.5: Plot of distribution function $G$ given by (2.13) for $n = 50$ and $m = 20, 55, 100$

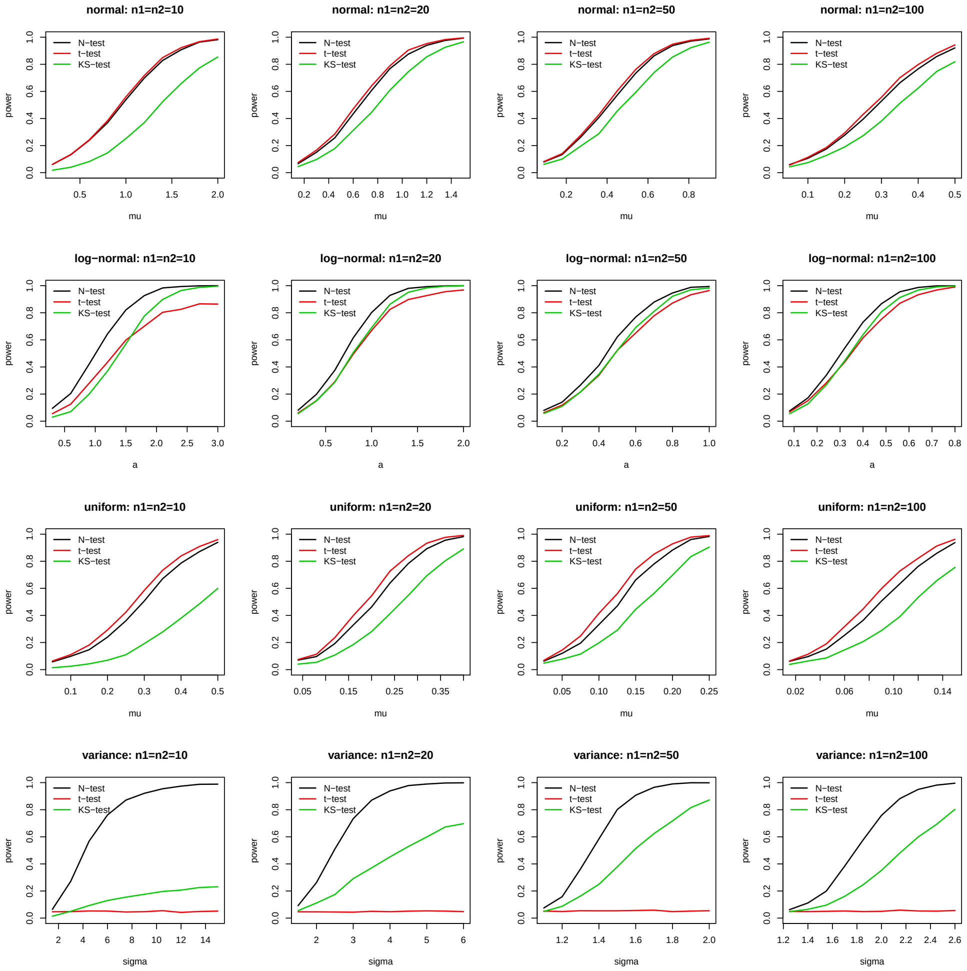

Figure 2.6: Comparison of power of $N$-test (black), $t$-test (red) and two-sample Kolmogorov-Smirnov test (green) for samples with sample sizes $n_1 = n_2 = 10, 20, 50, 100$ having distributions $N(0, 1)$ and $N(\mu, 1)$ (normal case); $\log N(0, 1)$ and $\log N(\mu, 1)$ (log-normal case); $U(0, 1)$ and $U(0, 1) + \mu$ (uniform case); $N(0, 1)$ and $N(0, \sigma^2)$ (variance case).

# Chapter 3

# Dependence of *p*-values of gene expression data

Gene expressions are highly correlated between genes. Therefore, marginal tests about genes are dependent as well as their *p*-values. In this chapter, we show how histograms of such *p*-values look like and how simple normalization such as proportion of gene expressions can change the structure of *p*-values. To do this we use *N*-test on HYPERDIP and TEL data. At the beginning, we consider the case when all hypotheses are true. Hence we use HYPERDIP and TEL data separately and we divide each of them into two halves. Thereafter, we consider cases when some hypotheses can be false. Hence we consider *p*-values of *N*-test between genes of HYPERDIP and TEL data. We show that this normalization has large impact on *p*-values. Results of this chapter were published in *Bubeliny* (2008).

## 3.1 All hypotheses are true

Firstly, we consider only HYPERDIP data. We divide these data into two halves and construct two samples for each gene consisted of 44 slides (we use up to 44 slides for the first sample and the remaining 44 for the second sample). For each gene, these samples are from the same distribution with distribution function $G_i^H$, $i = 1, \dots, 7084$. To emphasize the equality of distributions of these two samples for each gene $i$ (that is, the hypothesis about equality of distribution of $i$-th gene is true) we will write this hypothesis like $H_i' : G_i^H = G_i^H$.

If gene expressions were independent between themselves, the *p*-values for testing true hypotheses $H_i' : G_i^H = G_i^H$, $i = 1, \dots, 7084$ would have uniform distribution. In figure 3.1, we can see that the histogram of *p*-values for $H_i'$ has an obvious peak in the top about 0.85 and it is very different from the histogram of random variables having uniform distribution. It confirms that $p$-values are dependent.

Let us show how simple normalization, such as proportion of gene expressions, can

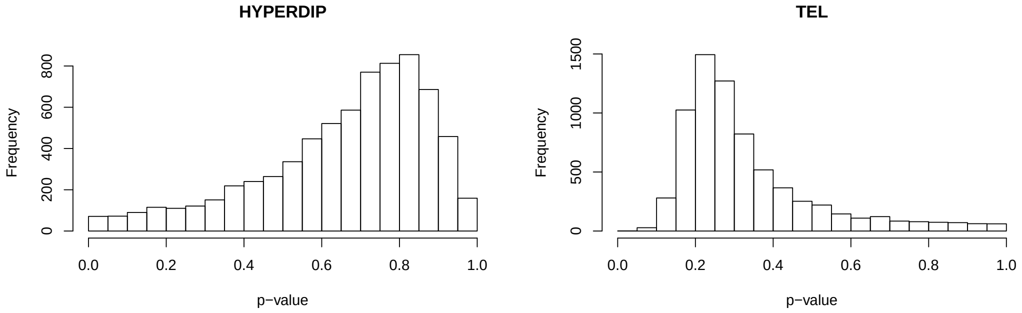

Figure 3.1: Histograms of $p$-values for true hypotheses for one kind of childhood leukemia. The left histogram is for HYPERDIP data, the right histogram is for TEL data.

help. We create random variables $\pi^1_{i,j} = x_{2i,j}/x_{2i-1,j}$, $j = 1, \dots, 44$ and $\pi^2_{i,k} = x_{2i,k}/x_{2i-1,k}$, $k = 45, \dots, 88$. For each fixed $i = 1, \dots, 3542$ these variables have the same distribution (denoted by $G^\pi_i$). Our goal is to test true hypotheses

$$H^\pi_i : \; G^\pi_i = G^\pi_i \text{ for each } i = 1, \dots, 3542,$$

simultaneously. From the histogram of $p$-values for these hypotheses (figure 3.2, the top left one) we can see that this histogram has different shape from previous one. There are almost equal columns and there is not a significant peak as it was in previous situation. Therefore, this histogram looks very similar to the histogram of sample from uniform distribution.

*Klebanov et al.* (2006) found out a new type of dependence, called type A dependence, which appears in microarray data. Let $x$ and $y$ be gene expression levels for gene $g_x$ and $g_y$, respectively. We say that pair $(g_x, g_y)$ is type A if $x$ and $y$ satisfy the condition $y = xz$, where $z$ is a positive random variable stochastically independent on $x$. Log$_2$ transformation of type A dependent random variables gives Y=X+Z, where $Y = \log_2 y$, $X = \log_2 x$ and $Z = \log_2 z$. According to independence of $x$ and $z$ we have that $\operatorname{Var} Y > \operatorname{Var} X$. Hence, this type of dependence is not symmetric.

The idea of type A dependence leads us to construct sorted (according to their variance) $\pi$ random variables. To distinguish unsorted and sorted $\pi$ random variables, we add the index $s$ to these variables. Therefore, we can define $\pi^{1s}_{i,j} = x_{(2i),j}/x_{(2i-1),j}$, $j = 1, \dots, 44$ and $\pi^{2s}_{i,k} = x_{(2i),k}/x_{(2i-1),k}$, $k = 45, \dots, 88$ and for fixed $i$ we denote the distribution function of $\pi^{1s}_{i,j}$ by $G^{\pi^s}_i$, which is the same as the distribution function of $\pi^{2s}_{i,j}$,

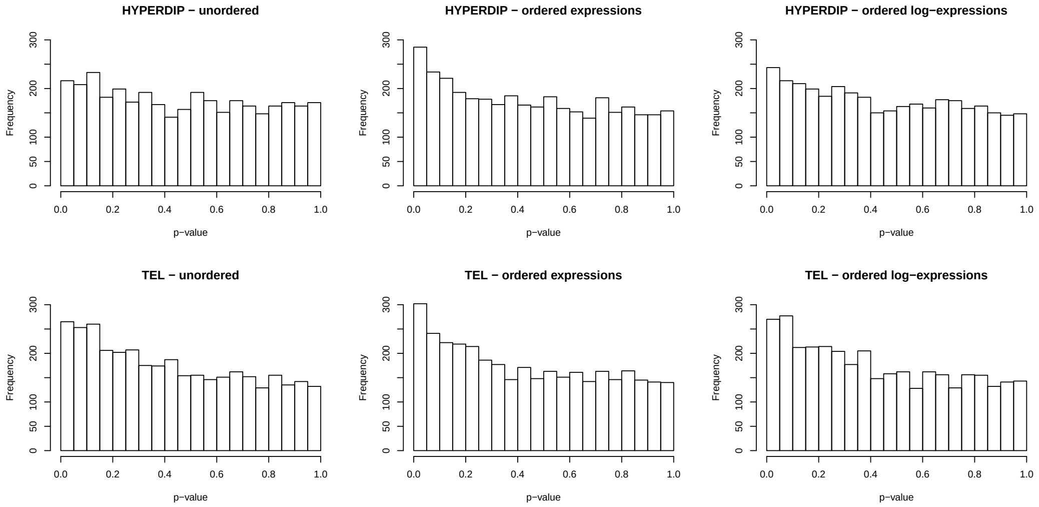

Figure 3.2: Histograms of $p$-values for true hypotheses of one kind of childhood leukemia for $\pi$ random variables (the top row for HYPERDIP data, the bottom row for TEL data). From the left to the right we consider hypotheses $H_i^\pi : G_i^\pi = G_i^\pi$ for unordered proportions of gene expressions, $H_i^{\pi^s} : G_i^{\pi^s} = G_i^{\pi^s}$ for ordered proportions of gene expressions according to variance of gene expressions and according to variance of gene $\log_2$-expressions.

where $x_{(k),j} = x_{l,j}$ and $l$ is the index of gene with $k$-th largest estimate of variance of expression levels. The histogram of $p$-values for hypotheses

$$H_i^{\pi^s} : G_i^{\pi^s} = G_i^{\pi^s} \quad i = 1, \ldots, 3542$$

is on the figure 3.2 (the top middle one). We observe a small change: $p$-values are much lower (columns at the beginning of the histogram are higher than for $p$-values greater than 0.2). However, this difference is not as dramatic as it was in previous comparison.

According to the type A dependence we can order genes by arranging them in increasing order of estimates of variances of gene $\log_2$-expressions. We create sorted $\pi$ random variables as in previous situation but with different ordering of genes. The histogram of $p$-values for hypotheses

$$H_i^{\pi^s} : G_i^{\pi^s} = G_i^{\pi^s}, \quad i = 1, \ldots, 3542$$

for this ordering is on the figure 3.2 (the top right). We can see that this histogram looks a little better than the previous one and it is very similar to the histogram for unordered data. Therefore, this ordering seems to be better than ordering by variance of gene expressions.

We can proceed the same way for TEL data as it was done for HYPERDIP data. We divide TEL data into 39 and 40 samples. Histogram for $p$-values of gene expressions for TEL data is given in figure 3.1. Histograms for all three situations for $\pi$ random variables are on the figure 3.2 (the bottom row). Histogram of $p$-values for gene expressions from TEL data has a peak about 0.25. All histograms of $p$-values for $\pi$ random variables of TEL data are similar to $\pi$ random variables of HYPERDIP data.

All previous histograms indicate that $\pi$ random variables are far less correlated than gene expressions. It proves that normalization could make gene expression data more workable.

## 3.2 Some hypotheses are false

So far, we only considered the situation where all testing hypotheses were true. Now we would like to know how the situation change if there are some false hypotheses. Therefore, we take HYPERDIP and TEL data for childhood leukemia together. We are interested in testing which genes are differentially expressed. It means that we would like to test hypotheses

$$H_i : \ G_i^H = G_i^T \text{ for each } i = 1, \ldots, 7084,$$

simultaneously. Histogram of $p$-values for these hypotheses is on the figure 3.3. We can see that there are 493 hypotheses with $p$-value less than or equal to 0.05. If we use Bonferroni inequality to decide which genes are differentially expressed at level $\alpha = 5\%$ we reject 111 hypotheses (critical value is $\frac{0.05}{7084}$).

It was shown in the previous section that $\pi$ random variables were far less dependent between genes than gene expressions of HYPERDIP data and gene expressions of TEL data. Therefore, as before we can define new $\pi$ random variables $\pi_{i,j}^H = x_{(2i),j}/x_{(2i-1),j}$ and $\pi_{i,l}^T = y_{(2i),l}/y_{(2i-1),l}$, $i = 1, \ldots, 3542$, $j = 1, \ldots, 88$ and $l = 1, \ldots, 79$. We can use unordered genes or we can order them as well. The problem is how we should sort the genes. There are some reasonable solutions. We can do it by arranging them in increasing order of estimates of variances of gene expressions in HYPERDIP data, in increasing order of estimates of variances of gene expressions in TEL data or in increasing order of estimates of variances of pooled HYPERDIP and TEL data. Because of type A dependence we can use all three ways of genes ordering according to estimates of variances of gene $\log_2$-expressions, too. We set all these options and estimate $p$-values for hypotheses

$$H_i^\pi : \ G_{(i)}^{\pi H} = G_{(i)}^{\pi T}, \quad i = 1, \ldots, 3542,$$

where $G_{(i)}^{\pi H}$ and $G_{(i)}^{\pi T}$ are distribution functions of $\pi_{(i)}$ random variables from HYPERDIP and TEL data, respectively, created from unordered genes or from one of considered

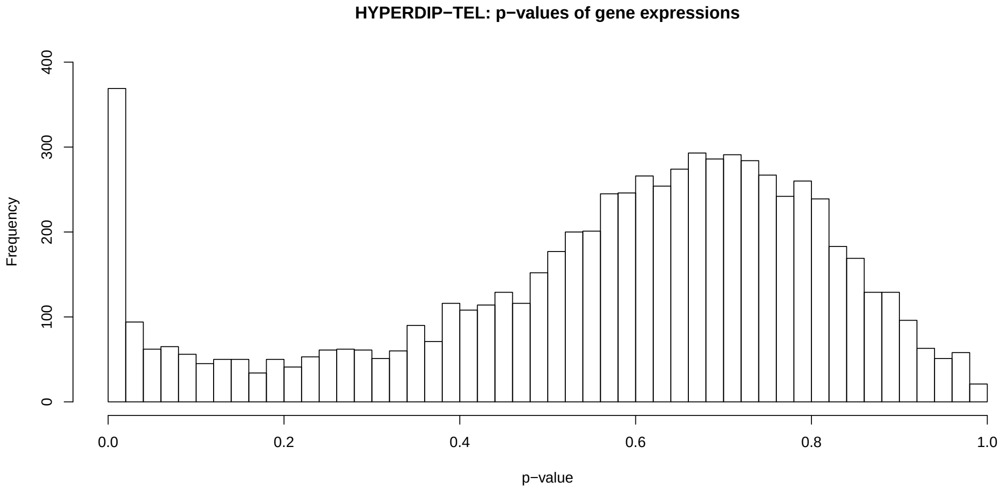

Figure 3.3: Histogram of $p$-values of testing equality of gene expressions between HYPERDIP and TEL data.

proposals of ordering of genes. Histogram of $p$-values for unordered genes is in figure 3.4. Histograms for ordered genes are very similar to those for unordered genes and they can be found in the supplement of this work. It can be surprising that there are a lot of $p$-values (much more than in the previous situation) less than or equal to 0.05 (it is about 43% of all). The number of rejected hypotheses for $\pi$ random variables according to Bonferroni inequality at significance level $\alpha = 5\%$ (critical value is $\frac{0.05}{3542}$) are in the table 1 (the top row).

One can say that there is a big difference in what we tested. In the first case,

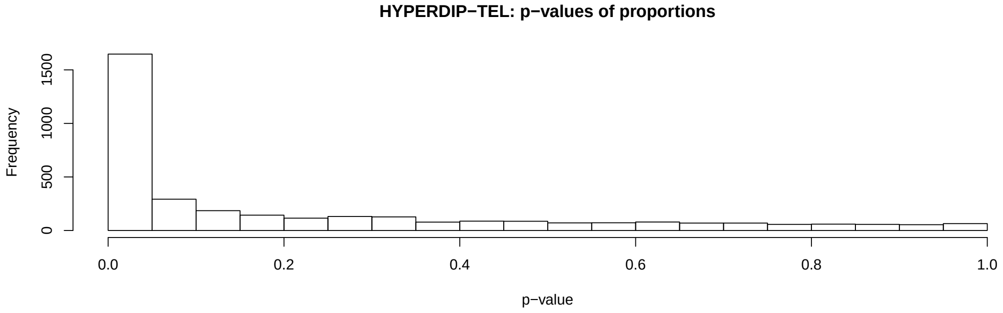

Figure 3.4: Histogram of $p$-values of proportions of gene expressions for unordered genes.

we tested 7084 hypotheses, one hypothesis for each gene, but in the second case we had 3542 hypotheses, one hypothesis for two genes. Because $N$-test is constructed for testing random vectors too, we can make 3542 non-overlapping pairs of genes and test if the joint distributions of these pairs of genes are the same for HYPERDIP data and TEL data. Therefore, we are interested in simultaneously testing hypotheses

$$H_i^2 : \ (G^H_{(2i-1)}, G^H_{(2i)}) \overset{D}{=} (G^T_{(2i-1)}, G^T_{(2i)}) \ \ i = 1, \dots, 3542,$$

where $G^H_{(j)}$ is the distribution function of gene expressions for j-th gene for HYPERDIP data and $G^T_{(j)}$ is the distribution function of gene expressions for j-th gene for TEL data. We consider 7 types of ordering as before. The first one is without ordering, three cases are obtained by arranging estimates of variances of gene expressions for HYPERDIP data, for TEL data and for pooled data in increasing order. The last three cases are obtained by arranging estimates of variances of $\log_2$-expressions for HYPERDIP data, for TEL data and for pooled data in increasing order. Histogram of $p$-values for hypotheses $H_i^2$ for pairs of unordered genes is in figure 3.5. Histograms corresponding to ordered pairs are similar and they can be found in the supplement of this work. We can see that this histogram is similar to the histogram for gene expressions from figure 3.3. The number of hypotheses we reject according to Bonferroni inequality are in table 3.1 (the bottom row). Their amount is far fewer than for $\pi$ random variables. Therefore, we can say that it is not the number of hypotheses (or how many genes we use in one hypothesis) but the manner of using gene expression levels for testing hypotheses that dramatically change the shape of histograms and the number of rejected hypotheses.

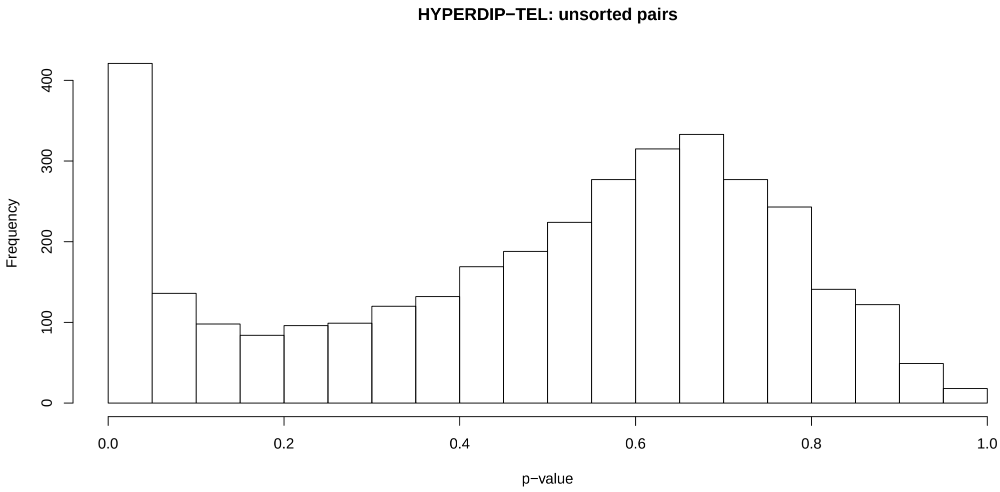

Figure 3.5: Histogram of $p$-values for pairs of gene expressions for unordered genes.

| order | unordered. | HYP | TEL | H-T | log-HYP | log-TEL | log-H-T |
|---|---|---|---|---|---|---|---|
| proportions | 626 | 665 | 751 | 656 | 604 | 643 | 595 |
| pairs | 94 | 80 | 91 | 78 | 73 | 88 | 62 |

Table 3.1: The number of rejected hypotheses according to Bonferroni inequality at significance level $\alpha = 5\%$ for all 7 types of ordering. The top row is for $\pi$ random variables, the bottom row is for pairs of gene expressions.

# Chapter 4

# Multiple testing procedures

Gene expression data usually consist of thousands of genes. Statistician working with such data often needs to test a lot of hypotheses simultaneously. Therefore, there is a need to use some multiple testing procedure that provides rejection regions for each of the hypotheses and guarantees controlling of predefined level $\alpha$. The goal of this chapter is to make an overview of different type I error rates, different types of power and multiple testing procedures. Many procedures rely on independence or special dependence structure of *p*-values or test statistics. For gene expressions, it is too difficult to verify such dependence due to strong correlation between genes. Therefore, in this chapter we introduce only such multiple testing procedures which control predefined level $\alpha$ for arbitrary test statistics joint distributions. A wider overview of multiple testing procedures is presented for example in *Dudoit et al.* (2003) or *Dudoit and van der Laan* (2008). At the end of this chapter we compare different multiple testing procedures.

## 4.1 Basic notes

Consider, that we want to test *M* hypotheses simultaneously. In any testing problem, two types of errors can be committed. The first type, called type I error or false positive, occurs if we reject true hypothesis. The second type, called type II error or false negative, occurs by non-rejecting false hypothesis. Ideally, we would like to minimize both of these errors. However, it is not possible. Therefore, we have to make some trade-off between these types of errors. Typically, this is done by minimizing type II error subject to type I error constraint. A multiple testing procedure specifies which hypotheses to reject, while controlling some type I error rate. Similar to single hypotheses testing, we can represent results of multiple testing procedures for each of *M* hypotheses in terms of confidence intervals for parameters of interest, rejection region for the test statistics and adjusted *p*-values. Adjusted *p*-values of multiple testing hypotheses are straightforward extensions of unadjusted *p*-values of single hypothesis testing. Adjusted *p*-value for $i$-th hypothesis, denoted by $\tilde{p}_i$, is the smallest nominal type I error level of multiple testing of $M$ hypotheses at which we reject this hypothesis. The hypothesis is rejected if adjusted $p$-value is lower or equal to the type I error rate $\alpha$. The smaller the adjusted $p$-value, the stronger the evidence to reject hypothesis.

Let $\mathcal{H}$ denote the set of true hypotheses and consider that the number of true hypotheses is $h$, that is $|\mathcal{H}| = h$. Likewise, let $\mathcal{A}$ denote the set of false hypotheses, then $|\mathcal{A}| = M - h$. Special case, when all hypotheses are true ($h = M$), is called complete null hypotheses. Usage of multiple testing procedure gives us the set of rejected hypotheses $\mathcal{R}$ and the set of non-rejected hypotheses $\mathcal{R}^c$. Then $\mathcal{R} \cap \mathcal{H}$ creates a set of false positives (type I errors) and $\mathcal{R}^c \cap \mathcal{A}$ a set of false negatives (type II errors). The situation is summarized in table 4.1, where

- the number of rejected hypotheses - $R = |\mathcal{R}|$,
- the number of false positives or type I errors - $V = |\mathcal{R} \cap \mathcal{H}|$,
- the number of false negatives or type II errors - $U = |\mathcal{R}^c \cap \mathcal{A}|$,
- the number of true negatives - $W = |\mathcal{R}^c \cap \mathcal{H}|$,
- the number of true positives - $S = |\mathcal{R} \cap \mathcal{A}|$.

Remark that $h$ and $a = M - h$ are unknown parameters, the number of rejected hypotheses $R$ is observable random variable and $S$, $U$, $V$ and $W$ are unobservable random variables.

| | Not rejected | Rejected | $\Sigma$ |
|---|---|---|---|
| True hypotheses | $W$ | $V$ | $h$ |
| False hypotheses | $U$ | $S$ | $a = M - h$ |
| $\Sigma$ | $M - R$ | $R$ | $M$ |

Table 4.1: Summary of different types of decisions and errors in multiple hypotheses testing.

We call a multiple testing procedure $\mathcal{M}$ monotone if for all vectors of $p$-values $\mathbf{p}$ and $\mathbf{p}'$ such that $\mathbf{p} \leq \mathbf{p}'$ ($p_i \leq p_i'$; $i = 1, \ldots, M$) the number of rejected hypotheses according to $\mathbf{p}$ is greater than or equal to the number of rejected hypotheses according to $\mathbf{p}'$, that is $|\mathcal{R}(\mathbf{p})| \geq |\mathcal{R}(\mathbf{p}')|$. The procedure is said to be cutting whenever the procedure rejects some hypotheses then they are those with the smallest $p$-values. Let $\mathcal{M}$ and $\mathcal{M}'$ be two multiple testing procedures. Following *Gordon* (2011), we say that a procedure $\mathcal{M}'$ dominates a procedure $\mathcal{M}$, if for any vector of $p$-values $\mathbf{p}$ we have $\mathcal{R}_{\mathcal{M}'}(\mathbf{p}) \supseteq \mathcal{R}_{\mathcal{M}}(\mathbf{p})$, that is $\mathcal{M}'$ rejects all hypotheses $H_i$ rejected by $\mathcal{M}$ (and maybe some others). In this case, we write $\mathcal{M}' \geq \mathcal{M}$. Let $C$ be a class of procedures and let $\mathcal{M} \in C$. We say that $\mathcal{M}$ is the most rejective (or optimal) in $C$ if $\mathcal{M} \geq \mathcal{M}'$ for all $\mathcal{M} \in C$. $\mathcal{M}$ is said to be unimprovable (or weakly optimal) in $C$ if the relations $\mathcal{M}' \in C$ and $\mathcal{M}' \geq \mathcal{M}$ imply that $\mathcal{M}' = \mathcal{M}$. Note that procedures $\mathcal{M}$ and $\mathcal{M}'$ may be incomparable, i.e., both relations $\mathcal{M}' \geq \mathcal{M}$ and $\mathcal{M} \geq \mathcal{M}'$ may be false. In particular, a class may contain more than one unimprovable multiple testing procedure. The most rejective multiple testing procedure in the class, if it exists, is unique.

When testing multiple hypotheses there are many definitions for type I error rates. The commonly used type I error rates are:

- family-wise error rate - $FWER = P(V > 0)$,
- generalized family-wise error rate - $gFWER(k) = P(V > k)$,
- per-comparison error rate - $PCER = \frac{E\,V}{M}$,
- per-family error rate - $PFER = E\,V$,
- false discovery rate - $FDR = E\frac{V}{R}$ (=0 if $R = 0$).

Notice, that $FDR$ can be rewritten as

$$FDR = E\frac{V}{max(R,1)} = E(\frac{V}{R}|R > 0)P(R > 0).$$

Therefore, $FDR \leq FWER$ and especially for completely null hypotheses all rejected hypotheses are type I error, $V/R = 1$ and therefore $FWER = FDR$. From Markov's inequality (A.2), we have

$$gFWER(k) = P(V \geq k+1) \leq \frac{1}{k+1}E\,V = \frac{1}{k+1}PFER$$

and specially for $k = 0$ we have $FWER \leq PFER$. Moreover,

$$PCER = \frac{E\,V}{M} = \frac{1}{M}\sum_{i=0}^{M} iP(V = i) \leq \sum_{i=1}^{M} P(V = i) = P(V > 0) = FWER.$$

Overall, we have $PCER \leq FWER \leq PFER$ and $FDR \leq FWER$. It means, that multiple testing procedure controlling $FWER$ generally results in fewer rejected hypotheses than multiple testing procedure controlling $FDR$ or $PCER$.

The most useful type I error rates for gene expression data are $FWER$ and $FDR$. Although inequality $FDR \leq FWER$ holds, each type I error means something different (has different interpretation) and we cannot say that one is better than the other. The correct choice of type I error rate depends on a situation. $FDR$ generally results in more rejected hypotheses and therefore in finding more true positives than $FWER$. On the other hand, if each type I error costs us a lot (money or human life), then it is better to use $FWER$-control instead of $FDR$-control.

Similarly, there are many definitions for power of multiple testing procedures, for example

- the probability of rejecting at least one false hypothesis - $AnyPwr = P(S > 0)$,
- the probability of rejecting all false hypotheses - $AllPwr = P(S = M - h)$,
- the average power - $AvgPwr = \frac{ES}{M-h}$,
- true discovery rate - $TDR = E\frac{S}{R}(= 0 \text{ if } R = 0)$.

Usually, there are two main types of multiple testing procedures, single-step and stepwise procedures. In single-step procedures, each hypothesis is tested using a rejection region, which is independent on the results of tests of other hypotheses. In stepwise procedures, the decision to reject a particular hypothesis depends on the results of the tests of other hypotheses. There are two main classes of stepwise procedures, step-down and step-up procedures. In step-down procedures, the most significant hypotheses are considered successively (for example in increasing order of their $p$-values). As soon as one hypothesis is not rejected, all less significant hypotheses are not rejected too. In step-up procedures, the least significant hypotheses are considered successively (for example in decreasing order of their $p$-values). As soon as one hypothesis is rejected, all more significant hypotheses are rejected too.

## 4.2 Multiple testing procedures for controlling $FWER$

### Bonferroni procedure

Perhaps the best known multiple testing procedure is Bonferroni procedure *Bonferroni* (1936). This procedure rejects any hypothesis with the unadjusted $p$-value less or equal to the cut-off $\tilde{\alpha} = \alpha/\mathrm{M}$. The set of rejected hypotheses is

$$\mathcal{R}(\alpha) = \{i : p_i \leq \frac{1}{M}\alpha\}.$$

The corresponding adjusted $p$-value for $i$-th hypothesis is $\tilde{p}_i = \min(Mp_i, 1)$, $i = 1, \ldots, M$.

**Theorem 4.2.1.** *Bonferroni procedure controls FWER at level $\alpha$ for arbitrary test statistics joint distributions, that is $P(V > 0) \leq \alpha$.*

*Proof.*

$$\begin{aligned} P(V > 0) &= P(\sum_{i\in\mathcal{H}} I(p_i \le \frac{1}{M}\alpha) > 0) \\ &= P(\bigcup_{i\in\mathcal{H}} \{p_i \le \frac{1}{M}\alpha\}) \\ &\le \sum_{i\in\mathcal{H}} P(p_i \le \frac{1}{M}\alpha) \\ &\le \frac{h}{M}\alpha \le \alpha, \end{aligned}$$

where the first inequality results from Bonferroni inequality (A.1) and the second inequality results from inequality (A.3). □

Although this procedure is generally considered as $FWER$ controlling multiple testing procedure, it controls mean number of false discoveries ($PFER$) at level $\alpha$ as well, because

$$\begin{aligned} E\,V &= E\sum_{i\in\mathcal{H}} I\{p_i \le \frac{\alpha}{M}\} \\ &= \sum_{i\in\mathcal{H}} P(p_i \le \frac{\alpha}{M}) \\ &= \frac{h}{M}\alpha \le \alpha. \end{aligned}$$

Bonferroni procedure is often considered as very conservative procedure. However, we can look at this procedure as a step-up procedure as well. According *Gordon* (2007) this procedure is unimprovable in the class of monotone step-up procedures controlling $FWER$.

## Holm procedure

Bonferroni procedure is simple to implement but it tends to be too conservative. Improvement in power can be achieved by step-down Holm procedure *Holm* (1979) which is step-down analogue of classical Bonferroni procedure. Without loss of generality, consider that the indexes $r_1, \ldots, r_M$ are such that $p_{r_1} \le \ldots \le p_{r_M}$. Then, the unadjusted $p$-values cut-offs for this procedure are $\tilde{\alpha}_{r_i} = \frac{1}{M-i+1}\alpha$, the set of rejected hypotheses is given by

$$\mathcal{R}(\alpha) = \{r_i : p_{r_l} \le \frac{1}{M-l+1}\alpha \;\; \forall l \le i\}$$

and the corresponding adjusted $p$-values are given by $\tilde{p}_{r_i} = \max_{k=1,\ldots,r_i}(\min\{(M-k+1)p_{r_k}, 1\})$, $i = 1, \ldots, M$.

**Theorem 4.2.2.** *Holm procedure controls $FWER$ at level $\alpha$ for arbitrary test statistics joint distributions, that is $P(V > 0) \leq \alpha$.*

*Proof.* Consider, that we have $h = |\mathcal{H}|$ true hypotheses. If Holm procedure rejects at least one true hypothesis, then

$$\min_{i \in \mathcal{H}} p_i \leq \alpha_{M-h+1} = \frac{1}{h}\alpha.$$

Thus, we have

$$\begin{aligned} P(V > 0) &\leq P(\min_{i \in \mathcal{H}} p_i \leq \alpha_{M-h+1}) \\ &= P(\bigcup_{i \in \mathcal{H}} \{p_i \leq \alpha_{M-h+1}\}) \\ &\leq \sum_{i \in \mathcal{H}} P(p_i \leq \alpha_{M-h+1}) \\ &\leq \sum_{i \in \mathcal{H}} \alpha_{M-h+1} \\ &= h\frac{1}{h}\alpha = \alpha, \end{aligned}$$

where the second inequality results from Bonferroni inequality (A.1) and the third inequality results from inequality (A.3). □

In *Gordon and Salzman* (2008), there was proved that Holm procedure dominates all monotone step-down procedures controlling $FWER$. The following example shows that the step-down condition cannot be removed.

**Example 4.2.1.** *Let $\mathcal{M}$ be a procedure with unadjusted p-values $p_1, ..., p_m$ which rejects all hypotheses, if $p_i \leq \alpha$ for all i , and accepts all hypotheses otherwise. This procedure is monotone and controls $FWER$ at level $\alpha$. Nevertheless, the relation $\mathcal{M} \geq$ Holm is not true: if $p_i = c$ ($\alpha/m < c < \alpha$), $i = 1, 2, \ldots, m$, then $\mathcal{M}$ rejects all hypotheses, while Holm procedure rejects none.*

Although Holm procedure does not dominates all monotone procedures controlling $FWER$, Gordon in *Gordon* (2011) showed that this procedure is unimprovable in the class of monotone multiple testing procedures controlling $FWER$.

## 4.2.1 Comparison of Bonferroni and Holm procedure

Bonferroni procedure is considered to be too conservative. However, we know that this procedure is unimprovable in the class of monotone step-up procedures controlling $FWER$. The step-down improvement of this procedure is Holm procedure. We know that this procedure is unimprovable in the class of monotone multiple testing procedures controlling $FWER$ and dominates all monotone step-down procedures controlling $FWER$. Hence, there cannot exist procedure which improves Holm procedure among procedures controlling $FWER$ at the same level. Holm procedure is less conservative than Bonferroni procedure. In the following simulation, we compared the difference between these two procedures.

We simulated (according to algorithm B.1) two independent samples of genes having multivariate normal distributions. We considered three different values of correlation coefficient $\rho$ equal to 0, 0.5, 0.9. We set the number of genes $m$ to be 300, 500 and 1000. The number of observations $n$ in each group was equal to 20 and 50. For the number of differentially expressed genes $k$ we considered $k = m/20, m/10, m/5$. We considered two alternatives for differentially expressed genes. In the first alternative, the mean value of differentially expressed genes was shifted about constant $C$ (we considered 30 equidistant values depending on setting of parameters). The mean value of differentially expressed genes in the second alternative was created by $k$-dimensional vector of i.i.d random variables having $N(C, 1)$ distribution. To compute unadjusted $p$-values for each gene we used $t$-test. According to Bonferroni procedure and Holm procedure we estimated the average power, $FWER$ and $PFER$ from 5000 repetitions.

Adjusted $p$-values of Bonferroni procedure are always less or equal to $p$-values of Holm procedure. Therefore, estimates of power, $FWER$ and $PFER$ of Holm procedure cannot be smaller than estimates of power, $FWER$ and $PFER$ of Bonferroni procedure. Results for $m = 500$, $n = 50$ and $\rho = 0.5$ are in figures 4.1 (alternative one) and 4.2 (alternative two). Complete results of this simulation are in the supplement of this work. We can see that the differences between estimates of $FWER$ and $PFER$ for Bonferroni procedure and Holm procedure are very small, especially for small $\mu$ and $k$. Hence, we can say that conservativeness of Bonferroni procedure is not a result of simplicity of this procedure but it comes from the principle of multiple testing procedures. Moreover, lines of estimates of average power for these two procedures are overlapped in all cases. Therefore, we will not suffer from lack of power if we use Bonferroni procedure for controlling $FWER$.

## HYPERDIP and TEL data

In case of simulations, we showed that results of Bonferroni and Holm procedure are almost equal. Now we compare these two procedures for $\log_2$ transformation of HYPERDIP and TEL data. If we compute adjusted $p$-values of $t$-test for according to these procedures we find out that there is just 116 (Bonferroni procedure) and 118 (Holm procedure) adjusted $p$-values are lower than 1 and 71 (for both cases) lower than 0.05. It confirms that both Bonferroni procedure and Holm procedure lead to very similar results in deciding which genes are differentially expressed and which are not.

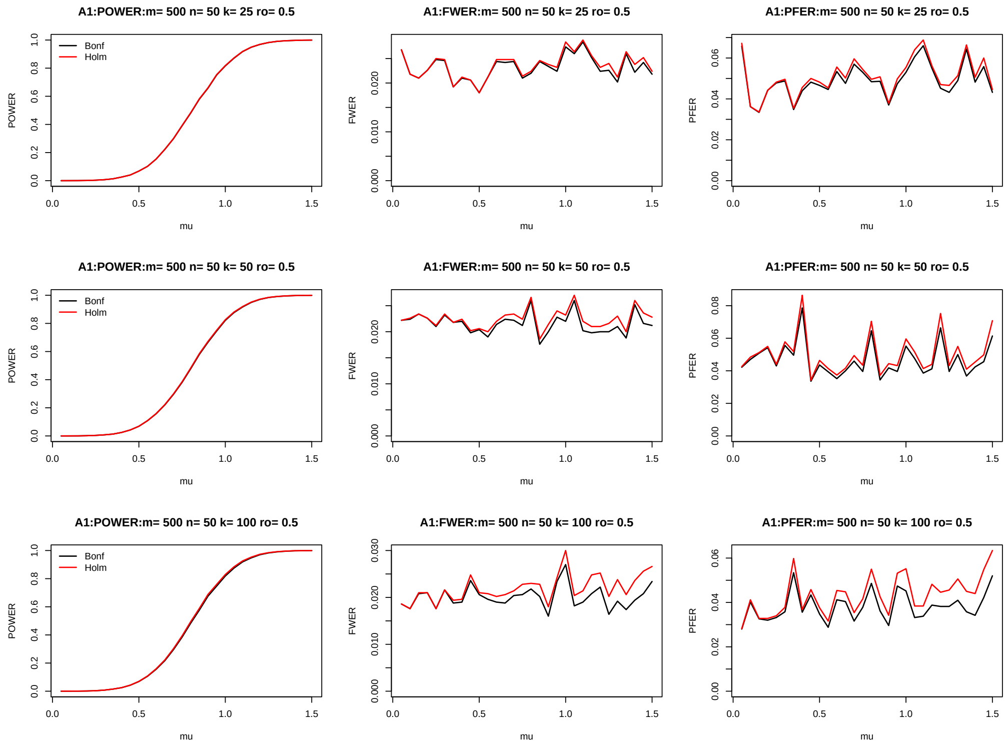

Figure 4.1: Comparison of Bonferroni and Holm procedure for alternative one and $m = 500$, $n = 50$ and $\rho = 0.5$.

## 4.3 Multiple testing procedures for controlling $gFWER$

### Lehmann-Romano procedures

In *van der Laan et al.* (2004), they showed that any $FWER$-controlling procedure can be straightforwardly augmented to control the $gFWER(k)$. But such procedures are too conservative.

In *Lehmann and Romano* (2005), they generalized Bonferroni procedure and Holm procedure to control generalized family-wise error rate $gFWER(k) = P(V > k)$. This generalization of Bonferroni procedure, called single-step Lehmann-Romano procedure, rejects any hypothesis with the unadjusted $p$-value less or equal to the cut-off $\tilde{\alpha} = \frac{k+1}{M}\alpha$. That is, the set of rejected hypotheses is given by

$$\mathcal{R}(\alpha, k) = \{i : p_i \leq \frac{k+1}{M}\alpha\}$$

and the corresponding adjusted $p$-values are thus given by $\tilde{p}_i \min(\frac{M}{k+1}p_i, 1)$, $i = 1, \dots, M$.

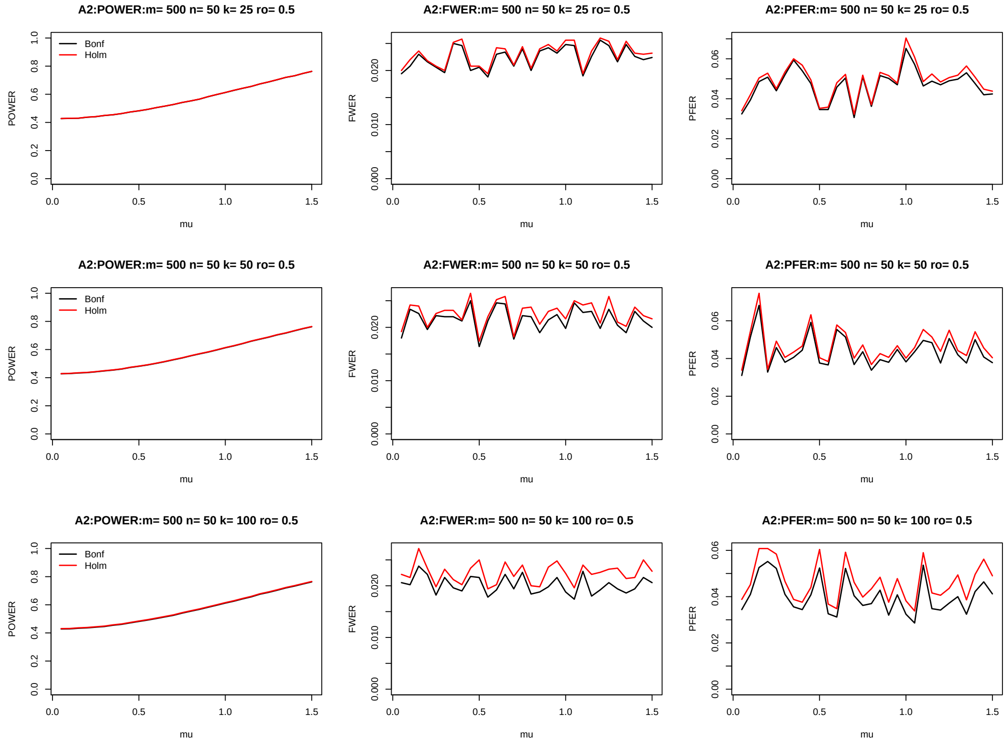

Figure 4.2: Comparison of Bonferroni and Holm procedure for alternative two and $m = 500$, $n = 50$ and $\rho = 0.5$.

**Theorem 4.3.1.** *Single-step Lehmann-Romano procedure controls the gFWER(k) at level $\alpha$ for arbitrary test statistics joint distributions, that is $P(V > 0) \leq \alpha$.*

*Proof.*

$$\begin{aligned}
P(V \geq (k+1)) &\leq \frac{1}{k+1} E\, V \\
&= \frac{1}{k+1} E\,(\sum_{i \in \mathcal{H}} I(p_i \leq \frac{k+1}{M}\alpha)) \\
&= \frac{1}{k+1} \sum_{i \in \mathcal{H}} P(p_i \leq \frac{k+1}{M}\alpha) \\
&\leq \frac{1}{k+1} \sum_{i \in \mathcal{H}} \frac{k+1}{M}\alpha \\
&= \frac{h}{M}\alpha \leq \alpha,
\end{aligned}$$

where the first inequality results from Markov's inequality (A.2) and the second inequality results from inequality (A.3). □

Note, that for $k = 0$ single-step Lehman-Romano procedure coincides with Bonferroni procedure.

The second Lehmann-Romano procedure, called step-down Lehmann-Romano procedure, is the generalization of Holm procedure. As for Holm procedure, without loss of generality consider, that the indexes $r_1, \ldots, r_M$ are such that $p_{r_1} \leq \ldots \leq p_{r_M}$. Then, the unadjusted $p$-values cut-offs for the step-down Lehmann-Romano procedure are given by

$$\tilde{\alpha}_{r_i} = \begin{cases} \frac{k+1}{M}\alpha & \text{if } i \leq k \\ \frac{k+1}{M+k+1-i}\alpha & \text{if } i > k, \end{cases}$$

the set of rejected hypotheses is given by

$$\mathcal{R}(\alpha, k) = \{r_i : p_{r_l} \leq \tilde{\alpha}_{r_l} \;\; \forall l \leq i\}$$

and the adjusted $p$-values are given by

$$\widetilde{p}_{r_i} = \begin{cases} \min\{\frac{M}{k+1}p_{r_i}, 1\} & \text{if } i \leq k \\ \max_{l=1,\ldots,i-k}\{\min\{\frac{M-l+1}{k+1}p_{r_{l+k}}, 1\}\} & \text{if } i > k. \end{cases}$$

**Theorem 4.3.2.** *Step-down Lehman-Romano procedure controls the gFWER($k$) at level $\alpha$ for arbitrary test statistics joint distributions, that is $P(V > k) \leq \alpha$.*

*Proof.* Consider, that we have $|\mathcal{H}| = h$ true hypotheses. For $h \leq k$, the probability of at least $k + 1$ false positives is equal to zero, that is $P(V > k) = 0$, so there is nothing to prove. Therefore, assume that $h > k$. Order $p$-values of true hypotheses and denote them by $q_1 \leq, \ldots, q_h$. Let $j$ be the index of $(k + 1)$-th ordered unadjusted $p$-value of true hypotheses, that is $p_j = q_{k+1}$. Then, the following inequalities hold $k + 1 \leq j \leq M - h + k + 1$ and we have $\tilde{\alpha}_j = \frac{k+1}{M+k+1-j}\alpha \leq \frac{k+1}{h}\alpha$. Hence,

$$\begin{aligned} P(V > k) &\leq P(p_j \leq \frac{k+1}{M+k+1-j}\alpha) \\ &\leq P(p_j \leq \frac{k+1}{h}\alpha) \\ &\leq \frac{k+1}{h}\alpha \leq \alpha, \end{aligned}$$

where the third inequality results from inequality (A.3) and the last inequality results from assumption that $h \geq k + 1$. □

Note, that for $k = 0$ step-down Lehman-Romano procedure coincides with Holm procedure.

## 4.4 Multiple testing procedures for controlling $FDR$

### Step-up Benjamini-Yekutieli procedure

In *Benjamini and Yekutieli* (2001), they proposed step-up Benjamini-Yekutieli procedure which controls false discovery rate $FDR = E\frac{V}{R}$ at level $\alpha$ for test statistics with arbitrary joint distribution. Without loss of generality consider, that indexes $r_1, \dots, r_M$ are such that $p_{r_1} \leq \dots \leq p_{r_M}$. Then, the unadjusted $p$-values cut-offs for the step-up Benjamini-Yekutieli procedure are $\tilde{\alpha}_{r_i} = \frac{i}{M C_M}\alpha$, $i = 1, \dots, M$, where $C_M = \sum_{i=1}^{M} \frac{1}{i}$. The set of rejected hypotheses is given by

$$\mathcal{R}(\alpha) = \{r_i : \exists l \geq i \text{ such that } p_{r_l} \leq \frac{l}{M C_M}\alpha\}.$$

The corresponding adjusted $p$-values are thus given by $\tilde{p}_{r_i} = \min_{h=i,\dots,M}\{\min(C_M \frac{M}{h} p_{r_h}, 1)\}$, $i = 1, \dots, M$.

**Theorem 4.4.1.** *Step-up Benjamini-Yekutieli procedure controls FDR at level $\alpha$ for arbitrary test statistics joint distributions, that is $E\frac{V}{max(R,1)} \leq \alpha$.*

*Proof.* Proof is not as straightforward as previous ones, therefore it is omitted. It can be found in section 4 of original work of *Benjamini and Yekutieli* (2001). □

In *Benjamini and Hochberg* (1995), they proved that the constant $C_M$ can be omitted for some special joint distributions of test statistics.

## 4.5 Empirical Bayes

An alternative way of dealing with multiple testing of hypotheses is considered in *Efron* (2003). It is based on empirical Bayes approach developed by Herbert Robbins in his paper *Robbins* (1964). This approach is closely related to $FDR$.

Let us consider to have test statistics $Y_i$, $i = 1, \dots, m$ for each of $m$ hypotheses. A very simple Bayesian model assumes that we have two classes of genes: differentially expressed genes (”different”) and non-differentially expressed genes (”non-different”) between two groups of observations. Let the prior probabilities for these two classes be $p_1$ and $p_0$ with corresponding prior densities $f_1(y)$ and $f_0(y)$ for statistic $Y$. Let $f(y)$ be a mixture density $f(y) = p_0 f_0(y) + p_1 f_1(y)$. From Bayes' theorem, we have the following posterior probabilities:

$$p_0(y) = P(\text{non-different}|Y = y) = \frac{P(Y = y|\text{non-different})P(\text{non-different})}{P(Y = y)} = \frac{p_0 f_0(y)}{f(y)}$$

and

$$p_1(y) = P(\text{different}|Y = y) = \frac{P(Y = y|\text{different})P(\text{different})}{P(Y = y)} = 1 - \frac{p_0 f_0(y)}{f(y)}. \qquad (4.1)$$

We conclude gene as differentially expressed, if its posterior probability $p_1(y)$ is greater than or equal to $1 - \alpha$.

Since there are too many unknown parameters we cannot calculate the exact value of $p_1(y)$. Therefore, it will be estimated by

$$\hat{p}_1(y) = 1 - \frac{\hat{p}_0 \hat{f}_0(y)}{\hat{f}(y)},$$

where $\hat{p}_0$, $\hat{f}_0(y)$ and $\hat{f}(y)$ are estimates of $p_0$, $f_0(y)$ and $f(y)$, respectively. Now we describe three different possibilities of estimating $p_0$, $f_0(y)$ and $f(y)$.

The first and the simplest proposal is considered in *Efron* (2003), where the density $f(y)$ is estimated by Poisson regression from histogram counts of $Y$-statistics. As an estimate of $f_0$ Efron took density of $Y$-statistics in case all hypotheses are true. For example, if we use $t$-test then $f_0$ will be estimated by density of $t$-distribution with corresponding degrees of freedom. Another unknown parameter is the probability of gene being non-different. As an estimate of this probability $p_0$ we take

$$\hat{p}_0 = \min_y\{\hat{f}(y)/\hat{f}_0(y)\},$$

which is the most conservative estimate, that makes all the posterior probabilities (4.1) nonnegative.

The second proposal is considered in *Efron* (2004). In his paper, Efron worked with $z$-values instead of the $Y$-statistics. Assume that for $i$-th hypothesis $i = 1, \ldots, m$ we have corresponding $p$-value $p_i$. Then $z$-value for this hypothesis is defined by $z_i = \Phi^{-1}(p_i)$, where $\Phi$ indicates the distribution function of standard normal distribution. If $i$-th hypothesis is true, then $z_i \sim N(0, 1)$. As previously, we can estimate mixture density $f(z) = p_0 f_0(z) + p_1 f_1(z)$ of $z$-values by Poisson regression from histogram counts of $z$-values and as $\hat{f}_0(z)$ we can take density of $N(0, 1)$.

In Efron's paper *Efron* (2004), there is also considered another estimate of $f_0$ (we will call it as the third proposal of empirical Bayes approach). It comes from the idea that $z$-values of non-different genes are concentrated in the peak of histogram of $z$-values. Therefore, as an estimate $\hat{f}_0$ Efron takes density of $N(\mu_0, \sigma_0^2)$, where

$$\mu_0 = \text{argmax}\{\hat{f}(z)\}$$

and

$$\sigma_0 = [-\frac{d^2}{dz^2}\log \hat{f}(z)]_{\mu_0}^{-\frac{1}{2}}.$$

In this work, we use Poisson regression (with polynomial of sixth degree) on histogram counts of 50 columns.

### 4.5.1 Comparison of empirical Bayes approaches

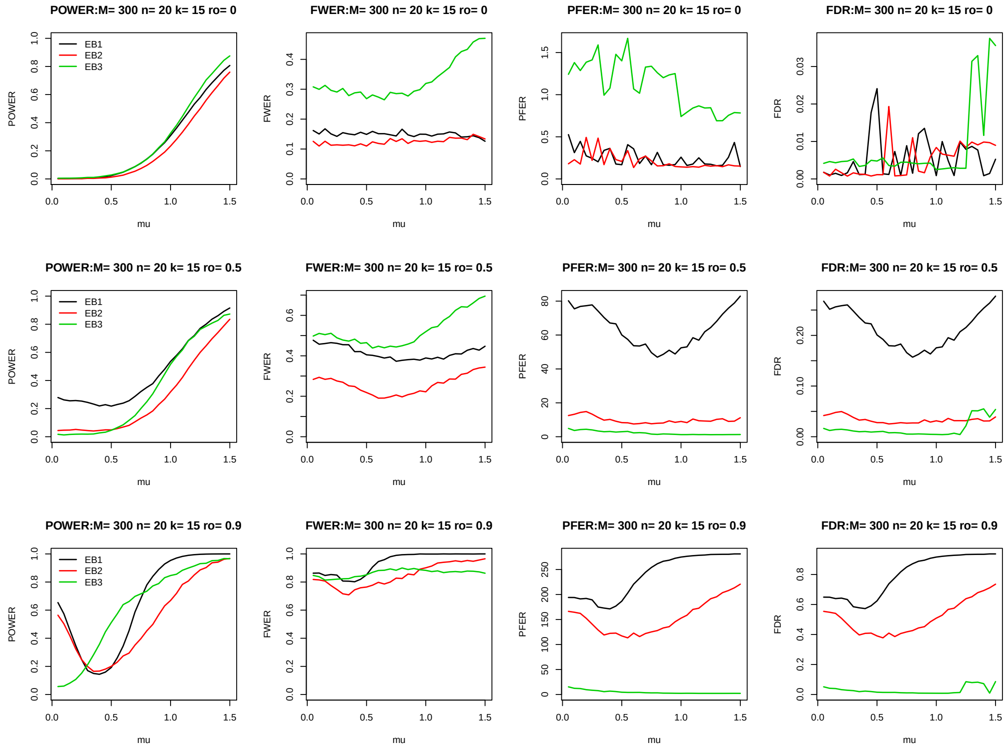

Figure 4.3: Estimate of Power, $FWER$, $PFER$ and $FDR$ by three empirical Bayes approaches for the number of genes $m = 300$, the number of diffrent genes $k = 15$, the number of observations in each group $n = 20$ and for correlation coefficient $\rho = 0, 0.5, 0.9$.

In order to compare three empirical Bayes approaches we performed simple simulation. We considered two independent samples of genes and we used these empirical Bayes approaches in order to decide which genes were differentially expressed between these groups. These two groups were created by two $m$-dimensional random samples having normal distribution $N_m(0, \Sigma)$ and $N_m(\mu_y, \Sigma)$, where $\mu_y$ had the first $m-k$ elements equal to 0 and the others $k$ elements equal to $\mu$, that is

$$\mu_y = (\underbrace{0, \ldots, 0}_{m-k}, \underbrace{\mu, \ldots, \mu}_{k})^T.$$

In this study, the parameter $\mu$ changed from 0.05 to 1.5 (by step 0.05). The covariance matrix $\Sigma$ was given by

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \dots & \rho & 1 & \rho \\ \rho & \dots & \dots & \rho & 1 \end{pmatrix},$$

where we set $\rho$ to be 0, 0.5 and 0.9. The data with such correlation structure were simulated according to algorithm (B.1). The sample sizes were set to $n = 20, 30, 50$, the number of genes was equal to $m = 300, 500, 1000$ and the number of different genes were $k = 15, 30, 60$. Each simulation was repeated 3000 times. We estimated the average power, $FWER$, $PFER$ and $FDR$ of considered three empirical Bayes approaches. Results of this simulation for $n = 20$, $m = 300$ and $k = 15$ are in figure 4.3. The whole results of this simulation can be found in the supplement of this work. For independent genes, the power of each empirical Bayes approach is similar. The second proposal has the most stable estimate of $PFER$ and $FDR$. Therefore, this proposal seems to be the best for independent genes. The results for dependent genes are different. Although the first proposal seems to have the best power, the estimates of $PFER$ and $FDR$ are very high what makes this proposal inapplicable. The third proposal holds $FDR$ for all setting and it has low estimate of $PFER$. Therefore, we should use the third empirical Bayes approach for dependent genes.

## HYPERDIP and TEL data

If we used three considered empirical Bayes approaches to real data, we would find that we reject almost all hypotheses. This fact is caused by very unreal estimates of $\hat{p}_0$. For $\log_2$ transformation of HYPERDIP and TEL data, these prior probabilities of gene not to be differentially expressed are estimated to be lower than 0.0001. Therefore, we use much realistic estimate of this probability. We fix $\hat{p}_0$ to be 0.9, 0.95 and 0.99. These setting cause some estimates of $p_1(y)$ to be negative. Hence we change $\hat{p}_1(y)$ to be $\hat{p}_1(y) = \max(0, 1 - \frac{\hat{p}_0 \hat{f}_0(y)}{\hat{f}(y)})$.

Now we can find which genes are differentially expressed according to empirical Bayes approach. The results are in table 4.2. We can see that the third approach of empirical Bayes has the best power (about three times greater than the other approaches). Moreover, the set of differentially expressed genes founded by the first approach is subset of differentially expressed genes founded by the second approach that is subset of set of differentially expressed genes founded by the third approach.

|  | EB1 | EB2 | EB3 |
|---|---|---|---|
| $\hat{p}_0 = 0.90$ | 85 | 112 | 320 |
| $\hat{p}_0 = 0.95$ | 81 | 111 | 317 |
| $\hat{p}_0 = 0.99$ | 80 | 110 | 311 |

Table 4.2: Number of rejected hypotheses for HYPERDIP and TEL data by empirical Bayes approach at significant level $\alpha = 0.05$.

## 4.6 Comparison of multiple testing procedures

*FDR* controlling procedures are generally considered to produce more true positives than *FWER* procedures. Therefore, we performed simple simulation in order to compare Bonferroni procedure, Benjamini-Yekutieli procedure and empirical Bayes approaches. In the following simulation, we did not use all three empirical Bayes approaches. According to previous results, we used the second proposal of empirical Bayes for independent genes and for dependent genes we used the third approach of empirical Bayes.

In this study, the data were simulated as follows. We considered two independent samples $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ having $m$-dimensional normal distribution $N_m(\mu_x, \Sigma)$ and $N_m(\mu_y, \Sigma)$. For simplicity, we considered $\mu_x = (0, \ldots, 0)^T$ and $\mu_y$ such that it had the first $m - k$ elements equal to 0 and the others $k$ elements equal to $\mu$. Moreover, we took equal covariance matrix for both samples with diagonal elements equal to one and non-diagonal elements equal to $\rho$. We set the number of genes $m$ to be 300, 500 and 1000. The correlation coefficient $\rho$ was set to 0, 0.5 and 0.9. The number of different genes $k$ was set to be 5%, 10% and 20% from the total number of genes. The shift parameter $\mu$ changed from 0.05 to 1.5 (by step 0.05). The number of observations in both samples were set to be 20 and 50. Each case was simulated 3000 times and we estimated average power, *FWER*, *PFER* and *FDR*.

In figure 4.4, there are results of simulation for correlation coefficient $\rho$ equal to zero for $m = 300$, $n = 20$ (the rest of results for $\rho = 0$ are in the supplement of this work). As we can see, that the second proposal of empirical Bayes has slightly greater power than Bonferroni and Benjamini-Yekutieli procedures, but it produces more false positives as well. Benjamini-Yekutieli procedure is more powerful than Benferroni procedure (except of small $\mu$, $m$ and $k$). Bonferroni procedures is the only one procedure, which controls *FWER* and *PFER* (and *FDR* as well), but empirical Bayes and Benjamini-Yekutieli controls *FDR* for each $\mu$ and they produce acceptable number of false positives. Results for correlated data and $m = 300$, $n = 20$ are in figure 4.5 and figure 4.6 (the complete results are in supplement of this work). They show that the third approach of empirical Bayes has far better power than other two considered procedures. Although empirical Bayes holds *FDR*, the number of false positives is too

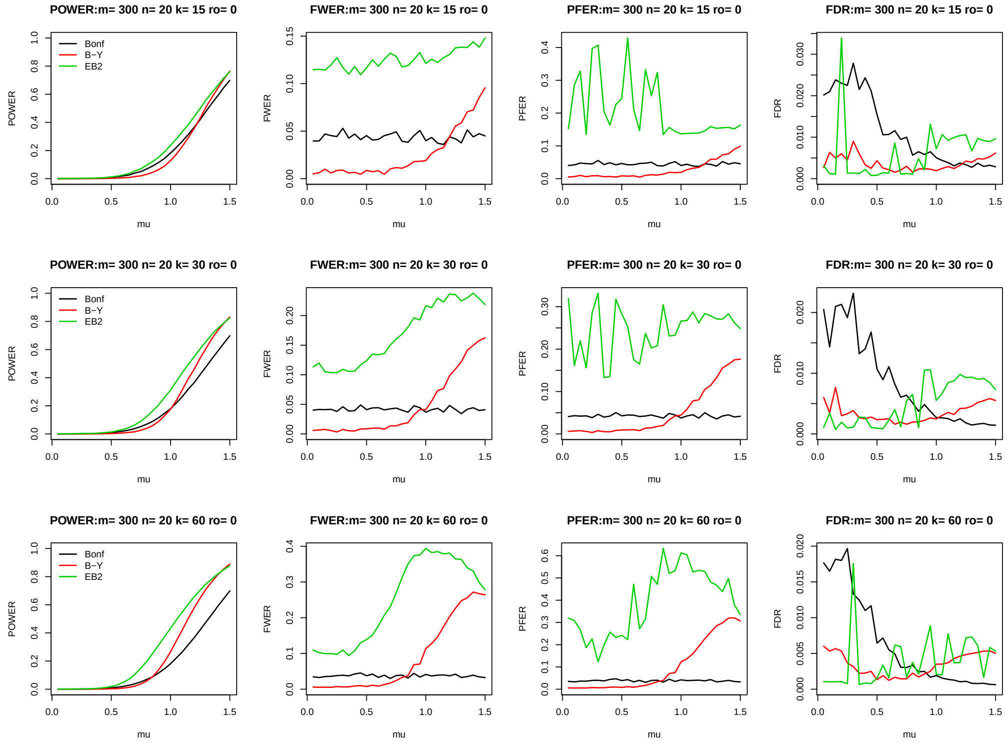

Figure 4.4: Estimate of Power, $FWER$, $PFER$ and $FDR$ by Bonferroni procedure, by Benjamini-Yekutieli procedure and by the second proposal of empirical Bayes approach for the number of genes $m = 300$, the number of observations in each group $n = 20$ and for correlation coefficient $\rho = 0$.

high in comparison with the other two procedures. Benjamini-Yekutieli procedure has better power than Bonferroni procedure. However, it seems to be too conservative for controlling $FDR$. Therefore, it is reasonable to use third approach of empirical Bayes instead of Benjamini-Yakutieli procedure for controlling $FDR$ in gene expression data.

## HYPERDIP and TEL data

Results of Bonferroni procedure and empirical Bayes approach for $\log_2$ transformation of HYPERDIP and TELL data were computed in previous sections of this chapter. However, results of Benjamini-Yekutieli procedure were not showed yet. At level $\alpha = 0.05$ this procedure discovered 92 genes to be differentially expressed. This is more than Bonferroni procedure, but more than 3 times less than third approach of empirical Bayes. Summary of results for these procedures are in table 4.3. Moreover, the set of

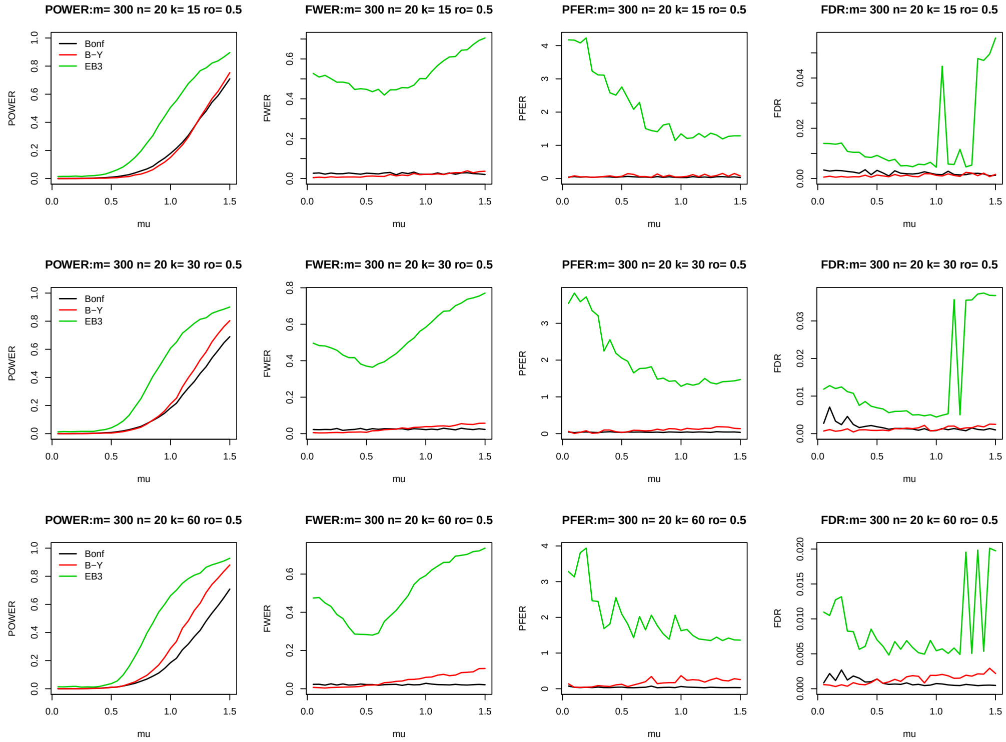

Figure 4.5: Estimate of Power, $FWER$, $PFER$ and $FDR$ by Bonferroni procedure, by Benjamini-Yekutieli procedure and by the third proposal of empirical Bayes approach for the number of genes $m = 300$, the number of observations in each group $n = 20$ and for correlation coefficient $\rho = 0.5$.

differentially expressed genes discovered by Bonferroni procedure is subset of differentially expressed genes discovered by Benjamini-Yakutieli procedure, which is subset of set of differentially expressed genes discovered by the third approach of empirical Bayes.

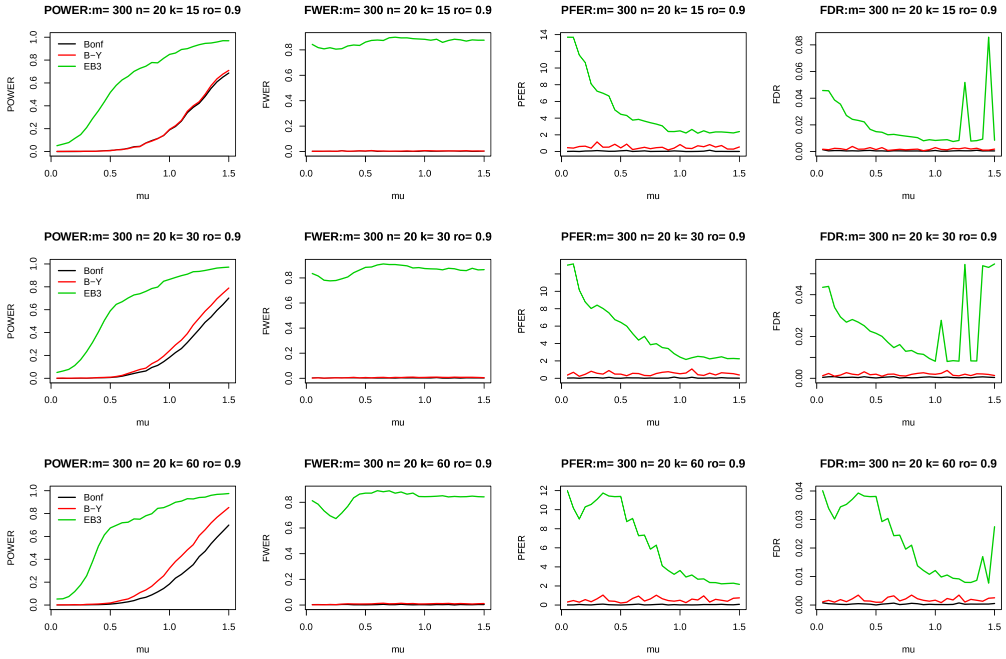

Figure 4.6: Estimate of Power, $FWER$, $PFER$ and $FDR$ by Bonferroni procedure, by Benjamini-Yekutieli procedure and by the third proposal of empirical Bayes approach for the number of genes $m = 300$, the number of observations in each group $n = 20$ and for correlation coefficient $\rho = 0.9$.

| procedure | Bonferroni | Empirical Bayes | Benjamini-Yekutieli |
|---|---|---|---|
| different genes | 71 | 317 | 92 |

Table 4.3: Number of rejected hypotheses for HYPERDIP and TEL data according to Bonferroni procedure, third empirical Bayes approach with $\hat{p}_0 = 0.95$ and Benjamini-Yekutieli procedure at significant level $\alpha = 0.05$.

# Chapter 5

# Normalizations

There are many sources of systematic variations in microarray experiments that affect measure of gene expressions. Common way of removing such variations is to normalize data (see e.g. *Yang et al.* (2002)). At the beginning of this chapter, we describe three types of normalizations: quantile normalization, global normalization and $\delta$-sequence. Moreover, we proposed some modification of $\delta$-sequence normalization. Thereafter, we show that although normalizations make data almost uncorrelated they change the gene expressions as well. Hence for deciding which genes are differentially expressed we have to use normalizations very carefully in order not to find too many false positives. Partial results of this chapter were published in *Bubeliny* (2013b).

## 5.1 Introduction

One of the problems of microarray data is that gene expressions are highly correlated between genes. Log$_2$ transformations of gene expressions are considered to have approximately normal distribution (see for example *Chen et al.* (2007)). In Figure 5.1, there are histograms of 100000 pairwise correlations of log$_2$-expressions between randomly chosen genes from HYPERDIP and TEL data. We can see that these correlations take values close to one (average correlation coefficient for HYPERDIP data is 0.91 and 0.92 for TEL data). This dependence of genes can influence many multiple testing procedures and the power of tests. Normalizations can be used to partially handle this problem. Testing of hypotheses is performed on these transformed (normalized) data. However, one can object to equality of testing with non-normalized data and with normalized data. Using normalized data, tests can break nominal level of multiple testing on which we would like to test hypotheses. It could bring many false positives, which we try to prevent.

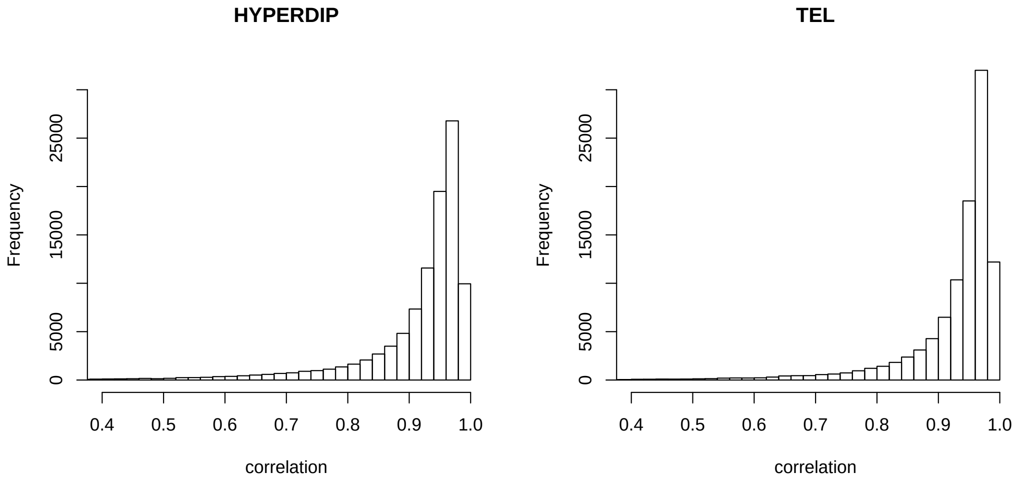

Figure 5.1: Histograms of 100000 estimates of random pairwise correlations of $\log_2$-gene expressions for HYPERDIP and for TEL data.

## 5.2 Normalizations

### Quantile normalization

The goal of the quantile normalization is to make the distribution of $\log_2$-gene expressions on each slide in a set of slides the same. The method is motivated by the idea of a $n$-dimensional quantile-quantile plot. In *Bolstad et al.* (2003), there was described algorithm for computing $X_{QN}$ (the matrix of $\log_2$-expressions after the quantile normalization). This algorithm is given as follows.

**Algorithm 5.1.**

1. *Given n slides of $\log_2$-gene expressions of length m, form a matrix X with m rows and n columns, where each slide is a column.*
2. *Sort each column of X to give $X_{sort}$.*
3. *Take means across rows of $X_{sort}$ and assign this mean to each element in the row to get $X'_{sort}$.*
4. *Obtain $X_{QN}$ by rearranging each column of $X'_{sort}$ to have the same ordering as the original matrix X.*

For two-sample problem, there are two possibilities how to use the quantile normalization. The first possibility is to use the quantile normalization separately for each sample. The second possibility is to create one pooled matrix of data and then make the quantile normalization on this pooled matrix. We will consider both of the mentioned possibilities of the quantile normalization.

## Global normalization

Global normalization of gene expression levels comes from the idea that gene expression is a product of two factors. The first factor is associated with gene production and the second factor is a constant unique for each slide. Therefore, we can imagine $\log_2$-expression $x_{ij}$ of gene $i$ from slide $j$ as a sum of two factors. The first factor depends on the gene $i$ and the second depends on the slide $j$, this is $x_{ij} = g_i + s_j$. Hence, it seems reasonable to subtract specific factor $S_j = s_j + c$ ($c$ is a constant independent on $i$ and $j$) from each $\log_2$-expression. There are two reasonable choices of $S_j$. The first one is slide mean, the second one is slide median. Thus, the algorithm for computing $X_{mean}$ (the matrix of $\log_2$-gene expressions after the global-mean normalization) and $X_{med}$ (the matrix of $\log_2$-gene expressions after the global-median normalization) is given as follows.

**Algorithm 5.2.**

1. *Given n slides of $\log_2$-gene expressions of length m, form a matrix X with m rows and n columns, where each slide is a column.*

2. *Take means or medians across the columns of X to obtain slide specific factors $\overline{X}_j$ or $X_j^{med}$ for $j = 1, \ldots, n$.*

3. *For each $j = 1, \ldots, n$ subtract from j-th column of X slide specific factor $\overline{X}_j$ (or $X_j^{med}$) to obtain $X_{mean}$ (or $X_{med}$).*

## $\delta$-sequence

In their paper, *Klebanov and Yakovlev* (2007) defined a new type of normalization. They eliminated slide effect in a different way as global normalizations do. Their $\delta$-sequence normalization is created by differences of non-overlapping $\log_2$-gene expressions. $\text{Log}_2$-gene expression data after $\delta$-sequence normalization can be defined as a $m^*$ ($m^* = \frac{m}{2}$, $m$ is tacitly assumed to be even) by $n$ dimensional matrix consisting of the random variables $\delta_{ij} = x_{2i-1,j} - x_{2i,j}$ $i = 1, \ldots, m^*$; $j = 1, \ldots, n$, where $x_{i,j}$'s are $\log_2$-gene expressions for $m$ genes from $n$ slides. They do not specify the order in which genes should be sorted. Therefore, in our simulation, we will consider two cases of ordering of gene expression data. The first case is a random permutation of genes. In the second case (monotonic), the ordering is according to $p$-values of gene expressions (we consider two-sample case) and in order to create $\delta_{ij}$ we pair $i$-th and ($m^* + i$)-th gene for $i = 1, \ldots, m^*$.

Another problem of $\delta$-sequence is that we examine difference of two genes. Therefore, we cannot make the decision for each gene separately. To decide which gene should be considered as a differentially expressed and which not, we propose four reasonable solutions.

Our first proposal $A$ not only pairs the $i$-th and the $(m^* + i)$-th gene, but pairs the $i$-th and $(m^* + i - 1)$-th gene (and the first and the last gene, respectively) for $i = 2, \dots, m^*$. If the $\delta$s created from some special gene are rejected both times then we will consider this gene as differentially expressed. If the $\delta$s calculated from some special gene are not rejected or are rejected just once we will consider this gene as non-differentially expressed.

If we consider monotonic ordering then each gene is paired with two similar (in $p$-value) genes. Therefore, some improvement for monotonic ordering can be achieved by computing the second $\delta$-sequence as the difference of the $i$-th and the $(m - i + 1)$-th gene for $i = 1, \dots, m^*$. Decisions for each gene are the same as in the previous case. We call this case as proposal $B$.

In the following proposal, there is tacitly assumed that there are at most $m^*$ false hypotheses. It seems reasonable to assume that the gene which evokes different expression of $\delta$-sequence is the one with lower $p$-value. Therefore, in our third proposal $C$ we assume that genes are ordered according their $p$-values. We pair $i$-th and $(m^* + i)$-th gene. The $i$-th $i = 1, \dots, m^*$ gene is said to be differentially expressed, if $\delta_{i.}$ is found to be differentially expressed.

Our fourth proposal $D$ is something like a step-down modification of the third case $C$. We consider the $i$-th gene for $i = 1, \dots, m^*$ as differentially expressed, if all $h$ hypotheses for $\delta_h$, $h = 1, \dots, i$ are rejected.

Now we explore how these normalizations change the structure of pairwise correlations of $\log_2$-expressions. In Figure 5.2, there are histograms of 100000 estimates of random pairwise correlations of $\log_2$-gene expressions after global-mean, global-median, quantile and random $\delta$-sequence normalization of HYPERDIP and TEL data. We can see, that these histograms are different from histograms of correlations of non-normalized $\log_2$-gene expressions, because they are symmetric and concentrated about zero.

## 5.3 Comparison of normalizations

### Simulations

In order to compare various normalizations, we performed the following simulation. We simulated two independent samples of $\log_2$-gene expressions $x_{ij}$ and $y_{ij}$, $i = 1, \dots, m$ (the number of genes), $j = 1, \dots, n$ (the number of slides) as random variables (highly

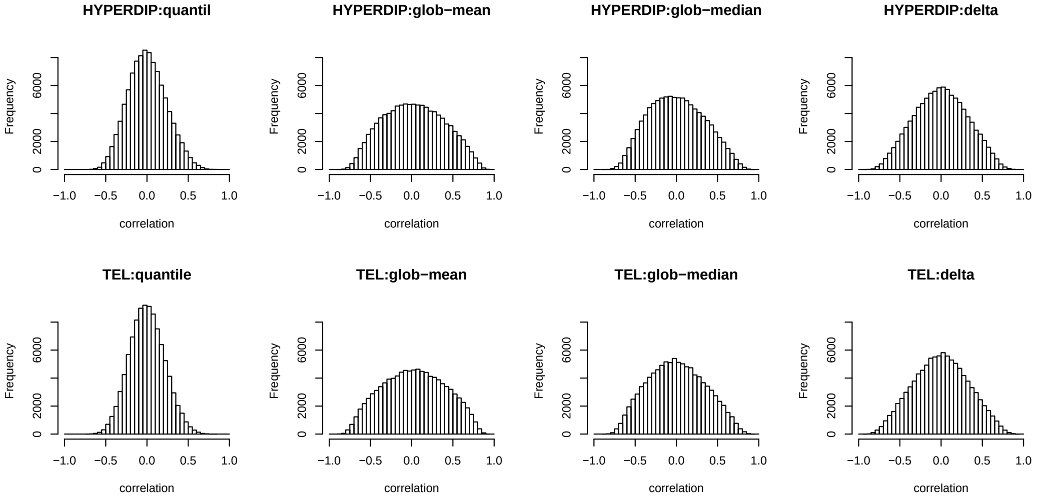

Figure 5.2: Histograms of 100000 estimates of random pairwise correlations of $\log_2$-gene expressions of HYPERDIP (upper row) and TEL (bottom row) after normalization. From left to right is Quantile-separate, Global-Mean, Global-Median and $\delta$-sequence normalization.

correlated in each slide) from normal distribution.

We generated $\log_2$-expressions for both samples by the algorithm (B.1). We set $\rho = 0.9$ and the number of genes equal to 500 (we consider only 500 genes because of computational complexity). We considered cases with equal number of slides for each stage $n = n_1 = n_2 = 10$, 25 and 50. In the second sample, we modified $k = 24$, 50, 90 and 200 genes (we create $k$ false hypotheses). We considered two different alternatives. In the first alternative, we shifted mean value of $k$ modified $\log_2$-gene expressions on each slide of the second sample about a constant $C$. For each setting $n$ and $k$ we considered 15 equidistant values of $C$. In the second alternative, we shifted $k$ $\log_2$-gene expressions on each slide of the second sample so that their expectations were created by random vector $\mu = (\mu_1, \ldots, \mu_k)$ with i.i.d. components having normal distribution $N(C, 1)$.

For each setting of $n$ and $k$, we performed 3000 simulations. For each simulated data we calculated quantile-separate, quantile-pooled, global mean, global median and proposals $A$-$D$ of $\delta$-sequence normalization. We studied average power (proportion of true positives and the number of differentially expressed genes $k$), mean number of false positive (estimate of $PFER$), probability at least one false positive (estimate of $FWER$) and estimate of false discovery rate (as relative frequency of present false positives among rejected hypotheses) according to Bonferroni procedure at nominal level $\alpha = 0.05$. We considered just $t$-test due to computation complexity and that genes differed in mean values (not in variance or distribution). The results for $N$-test are expected to be similar.

## Results

Due to the complexity of simulation we do not show all the results here. The complete results can be found in the supplement of this work.

If we look at results for non-normalized data we can see that $t$-test for such data holds nominal level $\alpha$ of $FWER$, but the power is very weak. On the other hand, there are some normalizations for which $t$-test roughly breaks level $\alpha$ of $FWER$. In figure 5.3 there are estimates of $FWER$ for data after $\delta$-sequence normalization for random and sorted pairs. Random pairs seem good, because $t$-test holds nominal level in all cases. In some cases (especially for alternative two) of sorted pairs, $t$-test after $\delta$-sequence normalization slightly breaks $\alpha$. However, some proposals of $\delta$-sequence are very bad. In table 5.1 there are some results for quantile normalization, proposal A and C of $\delta$-sequence. We can see that each of these normalizations is too risky in deciding which gene is differentially expressed and which is not. Therefore, we should not use them in deciding issue.

| | Alternative one | | | Alternative two | | |
|---|---|---|---|---|---|---|
| k | 24 | 50 | 90 | 24 | 50 | 90 |
| Quantile | 0.7600 | 0.7543 | 0.6927 | 0.7497 | 0.7503 | 0.7027 |
| Prop A-random | 0.8623 | 0.9753 | 0.3093 | 1.0000 | 1.0000 | 1.0000 |
| Prop A-sorted | 0.5990 | 0.8163 | 0.1140 | 1.0000 | 1.0000 | 1.0000 |
| Prop C | 0.1327 | 0.1677 | 0.1403 | 0.3207 | 0.3923 | 0.4227 |

Table 5.1: Estimate of $FWER$ for quantile normalization and proposals A and C of $\delta$-sequence for $n = 25$, $k = 24, 50, 90$ and $C = 0.4$.

In figure 5.4, there are estimates of level $\alpha$ of $FWER$ of $t$-test for non-normalized data after global mean and global median normalization, quantile pooled normalization and proposals B and D of $\delta$-sequence. Just for non-normalized data and proposal D, $t$-test holds level $\alpha$ in all cases. Global median is better than global mean normalization. However, for large number of different genes or large values of $C$ $t$-test for global median normalized data breaks nominal level $\alpha = 5\%$ as well. After quantile-pooled normalization $t$-test does not break nominal level of $FWER$ for small number of observations or small number of false hypotheses. Proposal B of $\delta$-sequence works well just for small number of differentially expressed genes. From this angle of view, there are just two possibilities how to normalize data. The first one is working with non-normalized data. The second reasonable normalization is proposal D of $\delta$-sequence.

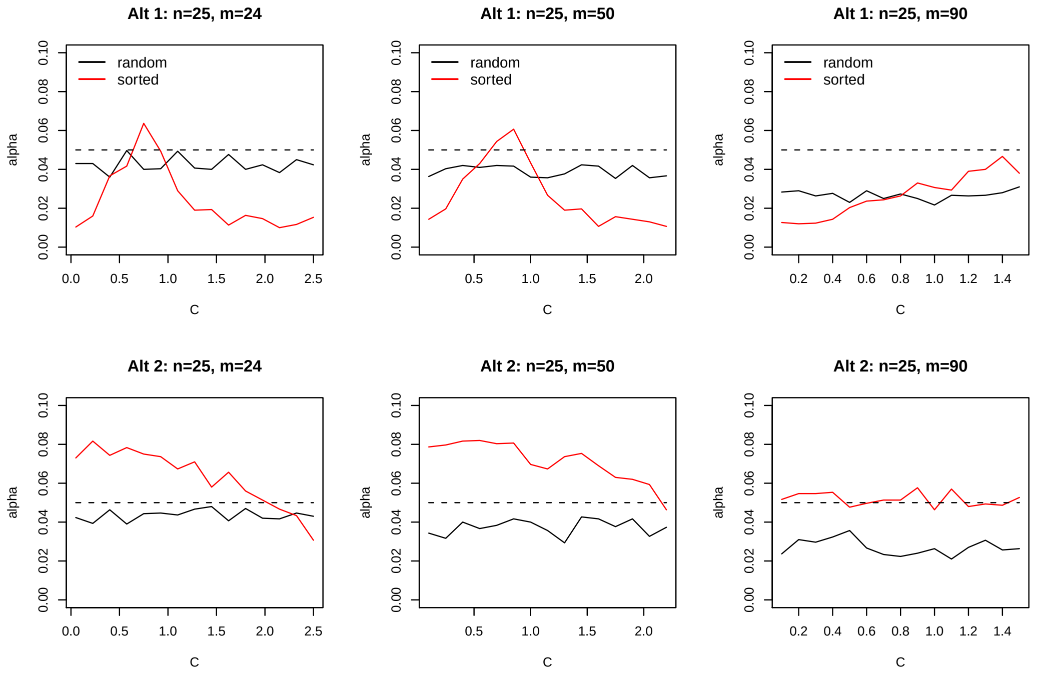

Figure 5.3: Plot of estimates of $FWER$ for $\delta$-sequence with random and sorted pairs for n=25.

Now we look at average power of $t$-test after these six normalizations. Some results are in figure 5.5. We can see that power of $t$-test after proposal D is better than for non-normalized data. The power after quantile-pooled normalization, according to another normalizations, decreases with increasing number of false hypotheses. The other three procedures have better power than proposal D. However, for small number of differentially expressed genes this difference is small and for large $k$ this difference is mainly influence by breaking nominal level $\alpha$ of $t$-test. Therefore, proposal D of $\delta$-sequence normalization is the best normalization for deciding which gene should be considered as a differentially expressed gene.

## 5.4 HYPERDIP and TEL data

Let us work with HYPERDIP and TEL data for childhood leukemia. For these data, we test which genes are differentially expressed. We would like to compare non-normalized testing with our best behaved fourth proposal $D$. If we applied classical approach with the $t$-test for non-normalized data, we would find 71 differentially expressed genes. For the $N$-test we discover 73 differentially expressed genes (67 genes are the same for $N$-

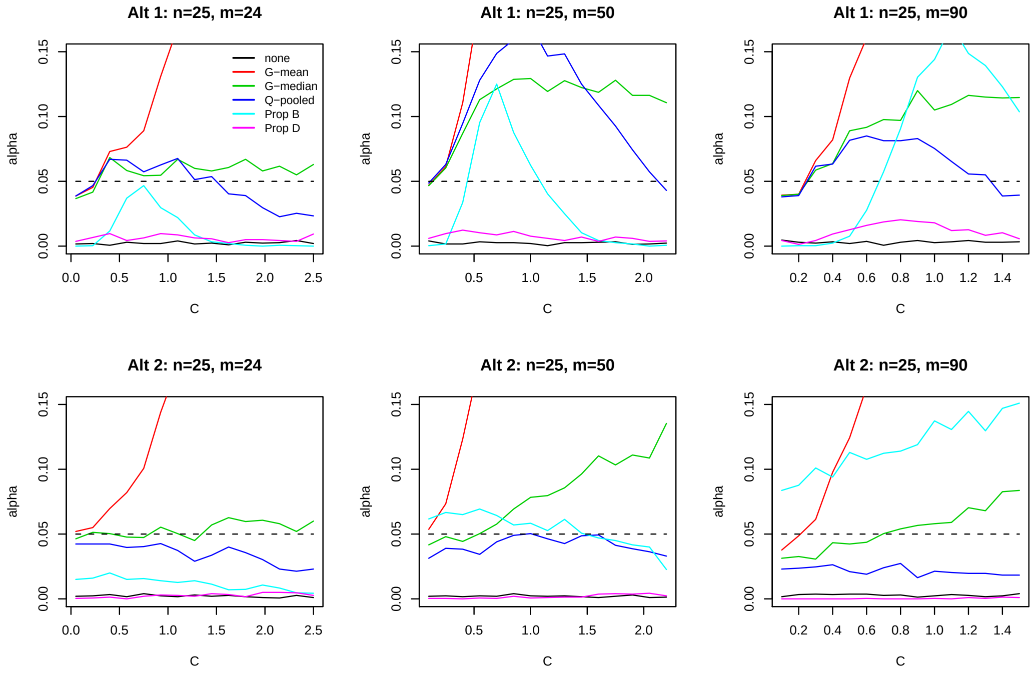

Figure 5.4: Plot of estimates of *FWER* for non-normalized data, global mean and median normalization,quantile-pooled normalization and proposal B and D of $\delta$-sequence normalization for n=25.

test and *t*-test). If we test according to proposal *D*, we have 81 differentially expressed genes by the *t*-test and 93 differentially expressed genes by the *N*-test (77 genes are the same). These results confirm that proposal *D* is an improvement of the classical approach using non-normalized data.

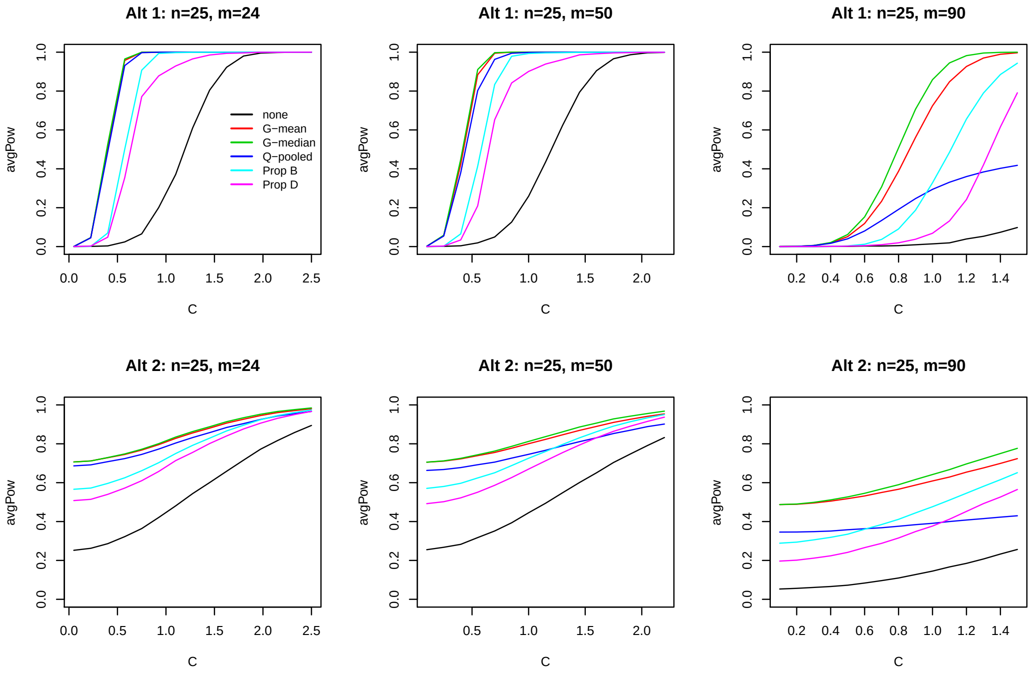

Figure 5.5: Plot of estimates of average power for non-normalized data, global mean and median normalization and for proposal B and D of $\delta$-sequence normalization for n=25.

# Chapter 6

# Gene sets

As we know, gene expressions are highly correlated between genes. Therefore, many papers work with gene sets (e.g. *Barry et al.* (2008)) instead of genes alone and therefore deal with multidimensional hypotheses. The most popular tests of two-sample problem for gene sets are Hotelling's test, *N*-test and tests derived from marginal *t*-statistics. We wrote about *N*-test in section 2.1. In what follows, we introduce tests based on marginal *t*-statistics and Hotelling's test. Our pre-study of Hotelling's test showed different behavior of this test in comparison to another considered tests. Therefore, for Hotelling's test, we derive some interesting properties in case of highly correlated data such as gene expressions are. Results of section about Hotteling's test were published in *Bubeliny* (2011). At the end of this chapter we compare the power of these tests.

## 6.1 Tests based on marginal *t*-statistics

There are two tests derived from marginal *t*-statistics which are often used for gene sets. The first one is based on sum of squares of marginal *t*-statistics and the second is based on sum of absolute values of marginal *t*-statistics, that is

$$T_{sq} = \sum_{i=1}^{m} t_i^2$$

and

$$T_{abs} = \sum_{i=1}^{m} |t_i|,$$

where $t_i$ is marginal *t*-statistic for *i*-th gene from gene set consists of *m* genes. The critical values of these statistics are not known. Therefore, we estimate the *p*-values of these tests by permutations of slides.

## 6.2 Hotelling's test

Hotelling's test is a multidimensional extension of $t$-test. Similar to $t$-test, we can consider both one-sample and two-sample Hotelling's test. One-sample case deals with hypothesis that the expected value of a sample from multidimensional normal distribution is equal to some given vector. In the two-sample case, it deals with the hypothesis of the equality of expected values of two samples from multidimensional normal distributions (with equal covariance structure). We will focus on the two-sample Hotelling's test.

Suppose we have two independent samples (of sizes $n_x$ and $n_y$, respectively) from two $m$-dimensional normal distributions with identical covariance matrices equal to $\Sigma$. In other words, we consider $X_1, \dots, X_{n_x}$ as i.i.d. random vectors having $N_m(\mu_x, \Sigma)$ and $Y_1, \dots, Y_{n_y}$ as i.i.d. random vectors having $N_m(\mu_y, \Sigma)$ ($X_i$ and $Y_j$ are independent for all $i = 1, \dots, n_x$; $j = 1, \dots, n_y$). For simplicity we assume that $m < n_x + n_y - 1$. Our goal is to test the hypothesis $H : \mu_x = \mu_y$ against the alternative $A : \mu_x \neq \mu_y$. Hotelling's test for this hypothesis is based on the statistic

$$T^2 = \frac{n_x n_y}{n_x + n_y} (\bar{X} - \bar{Y})^T S^{-1} (\bar{X} - \bar{Y}), \tag{6.1}$$

where

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

$$\bar{Y} = \frac{1}{n_y} \sum_{i=1}^{n_y} Y_i$$

and

$$S = \frac{\sum_{i=1}^{n_x} (X_i - \bar{X})(X_i - \bar{X})^T + \sum_{i=1}^{n_y} (Y_i - \bar{Y})(Y_i - \bar{Y})^T}{n_x + n_y - 2}.$$

$T^2$ is related to the $F$-distribution by

$$\frac{n_x + n_y - m - 1}{m(n_x + n_y - 2)} T^2 \sim F(m, n_x + n_y - m - 1). \tag{6.2}$$

For more details about Hotelling's test see e.g. *Chatfield and Collins* (1980). We made the assumption $m < n_x + n_y - 1$ for two reasons. For $m \geq n_x + n_y - 1$ the estimate $S$ of $\Sigma$ results in an irregular matrix, so that $S^{-1}$ does not exist and moreover numerator of (6.2) is non-positive as well as the degree of freedom of the $F$-distribution. Therefore, in such cases we use pseudo-inversion of $S$ and in order to estimate $p$-value of $H$, we use permutations of vector $(X_1, \dots, X_{n_x}, Y_1, \dots, Y_{n_y})$.

### 6.2.1 Hotelling's test for dependent data

Consider that we have two independent multidimensional samples from normal distribution. We would like to test hypothesis suggesting the equality of expected values in these two samples. Assume for simplicity that all elements on the main diagonal of the covariance matrix $\Sigma$ for both samples are equal to 1 and all other elements are equal to $\rho > 0$, i.e.

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \dots & \rho & 1 & \rho \\ \rho & \dots & \dots & \rho & 1 \end{pmatrix}.$$

Further on, we assume that $\mu_x = (0, \dots, 0)^T$, but $\mu_y$ has first $k$ elements equal to 1 and the others equal to 0, i.e.

$$\mu_y = (\underbrace{1, \dots, 1}_{k}, \underbrace{0, \dots, 0}_{m-k})^T.$$

For large $n_x$ and $n_y$, the matrix $\Sigma$ and its estimate $S$ are approximately the same as well as the differences between the expected values $(\mu_x - \mu_y)$ and between the mean values $(\overline{X} - \overline{Y})$. When dealing with real data, $n_x$ and $n_y$ might not be large enough, but for easier insight to the problem we use the approximations $S \approx \Sigma$ and $\overline{X} - \overline{Y} \approx \mu_x - \mu_y$. In this case $S^{-1} \approx \Sigma^{-1}$, that is

$$S^{-1} \approx \Sigma^{-1} = \begin{pmatrix} \omega & -\beta & -\beta & \dots & -\beta \\ -\beta & \omega & -\beta & \dots & -\beta \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\beta & \dots & -\beta & \omega & -\beta \\ -\beta & \dots & \dots & -\beta & \omega \end{pmatrix},$$

where $\omega = \frac{(1+(m-2)\rho)}{(1-\rho)(1+(m-1)\rho)}$ and $\beta = \frac{\rho}{(1-\rho)(1+(m-1)\rho)}$. For fixed $n_x$ and $n_y$ we can consider the fraction $\frac{n_x n_y}{n_x + n_y} = c$ of Hotelling's statistic (6.1) as a normalizing constant. Let us denote by $T^{*2}$ Hotelling's statistic with $\Sigma^{-1}$ instead of $S^{-1}$ and $\mu_x - \mu_y$ instead of $\overline{X} - \overline{Y}$ divided by the constant $c$. Then $T^{*2}$ is squared Mahalanobis distance of $\mu_x$ and $\mu_y$ and it is given by

$$T^2/c \approx T^{*2} = (\mu_x - \mu_y)^T \Sigma^{-1} (\mu_x - \mu_y)$$

$$= (\underbrace{1,\ldots,1}_{k},\underbrace{0,\ldots,0}_{m-k})\begin{pmatrix} \omega & -\beta & -\beta & \ldots & -\beta \\ -\beta & \omega & -\beta & \ldots & -\beta \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\beta & \ldots & -\beta & \omega & -\beta \\ -\beta & \ldots & \ldots & -\beta & \omega \end{pmatrix}\begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= k\omega - (k^2 - k)\beta = \frac{k(1+(m-2)\rho) - k(k-1)\rho}{(1-\rho)(1+(m-1)\rho)} = \frac{k(1+(m-k-1)\rho)}{(1-\rho)(1+(m-1)\rho)}. \qquad (6.3)$$

Let us note that it does not matter if $\mu_y$ consists of ones and zeros or equals to a constant $a$ and zeros. In the latter case, squared distance $T^{*2}$ would be multiplied by $a^2$. Now we will work with $T^{*2}$ and investigate its behavior.

If we changed $k$ to $k+1$ (meaning that we add one more different marginal distribution) we would expect that $T^{*2}$ increases as well as the power of Hotelling's test does. For better understanding let the number of ones in $\mu_y$ be the index of $T^{*2}$ (we will write it only when it is needed). Now we change $k$ to $k+1 = h$ and we have

$$T^{*2}_{k+1} = T^{*2}_{k} + \omega - 2k\beta.$$

If we expected that $T^{*2}$ is an increasing function of $k$ then $\omega - 2k\beta$ should be greater then zero. However, we have

$$\omega - 2k\beta = \frac{1+(m-2)\rho}{(1-\rho)(1+(m-1)\rho)} - \frac{2k\rho}{(1-\rho)(1+(m-1)\rho)} = \frac{1+(m-2k-2)\rho}{(1-\rho)(1+(m-1)\rho)}.$$

Since the denominator is greater than zero, then $\omega - 2k\beta > 0$ if and only if $\frac{1}{2k+2-m} = \frac{1}{2h-m} > \rho$. It means that for not very small values of $\rho$'s and $k > \frac{m}{2} - 1$ the square Mahalanobis distance $T^{*2}$ is a decreasing function of $k$. This means that maximal power of Hotelling's test (as a function of $k$) is not always attained for $k = m$ but for $\rho$'s which are not very small we have maximal power for $k$ near $\frac{m}{2}$. Some examples of the behavior of $T^{*2}$ as a function of $k$ are illustrated in figure 6.1.

However, this issue is not the only one that is surprising about Hotelling's test. Now we look if $T^{*2}_1$ is always lower than $T^{*2}_m$. It is the case when one different marginal distribution influences more than all $m$ different distributions. Therefore, we need to compare $\omega$ with $m\omega - m(m-1)\beta$. We have

$$T^{*2}_1 - T^{*2}_m = \omega - m\omega + m(m-1)\beta = (m-1)\frac{(1-2\rho)}{(1-\rho)(1+(m-1)\rho)}.$$

It means that $T^{*2}_1 - T^{*2}_m < 0$ if and only if $\rho < 0.5$. Therefore, we can say that for $\rho > 0.5$ Hotelling's test has better power for alternative with only one marginal shift than for alternative that all marginal distributions are equally shifted. It can be seen from figure

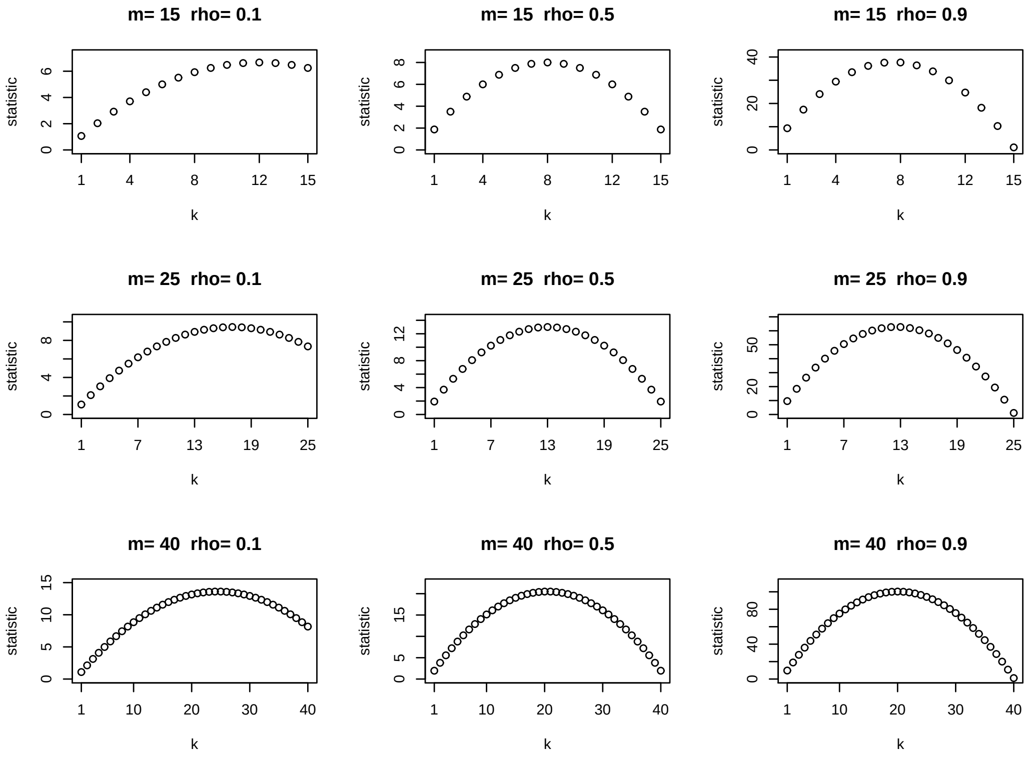

Figure 6.1: Plots of $T^{*2}$ for $m = 15, 25, 40$; $\rho = 0.1, 0.5, 0.9$; and $k = 1, \dots, m$. Notice: each plot is differently scaled!

6.1 as well. Moreover, $T^{*2}$ is an increasing function of $\rho$, that may seem to be surprising as well.

Let generalize expected value $\mu_y$ to have components $(a_1, \dots, a_m)$. We are interested in for which $\mu_y \in R^m$ the squared Mahalanobis distance has the same value. For some $d > 0$ we define the set

$$E_d = \{\mu_y = (a_1, \dots, a_m); \mu_y^T \Sigma^{-1} \mu_y = d^2\}.$$

This set is created by iso-distance curves, i.e. ellipsoids with center in $(0, \dots, 0)$. Let denote the eigenvalues of matrix $\Sigma^{-1}$ by $\lambda_1, \dots, \lambda_m$ and the eigenvectors corresponding to these eigenvalues by $\gamma_1, \dots, \gamma_m$. Then the principal axes of $E_d$ are in the direction of $\gamma_i$; $i = 1, \dots, m$ and the half-lengths of the axes are given by $\sqrt{\frac{d^2}{\lambda_i}}$; $i = 1, \dots, m$. In our case with $\Sigma^{-1}$, the eigenvalues $\lambda_1 = \lambda_2 = \dots = \lambda_{m-1} = \frac{1}{1-\rho}$ and $\lambda_m = \frac{1}{1+(m-1)\rho}$. The eigenvector corresponding to the smallest eigenvalue $\lambda_m$ is equal to $\gamma_m = (1, \dots, 1)$.

Therefore, squared Mahalanobis distance has the slowest increase in this direction.

### 6.2.2 Two-dimensional data

Let us look at Hotelling's test in the two-dimensional case. Some plots of two-dimensional ellipsoids for different values of the correlation coefficient $\rho$ are given in figure 6.2. The squared Mahalanobis distance has the weakest increase in the direction of $a_1 = a_2$, while the fastest increases is observed towards the direction of $a_1 = -a_2$. For example, for $\rho = 0.9$ and $d = \omega$ the principal axes are equal to 3.162 and 0.725. It means that for $a_1 = a_2 = \sqrt{\frac{3.162^2}{2}} = 2.236$ squared Mahalanobis distance is the same as for $a_1 = 1$, $a_2 = 0$ (or for $a_1 = -a_2 = \sqrt{\frac{0.725^2}{2}} = 0.513$ as well). Hence, if there is only one marginal distribution shifted by one unit, then the power of Hotelling's test is expected to be the same as if both marginal distributions were equally shifted (in the same direction) by 2.236 units (for the shift in opposite direction it should be only 0.513 unit). These results are in contradiction with other multidimensional tests. For example, consider some test based on marginal $t$-statistics. The power of this test is higher if both distributions are shifted by the same amount (both $t$-statistics are ”large”, not depending on direction of shift) than if there is only one marginal distribution shifted (one $t$-statistic is ”near” zero).

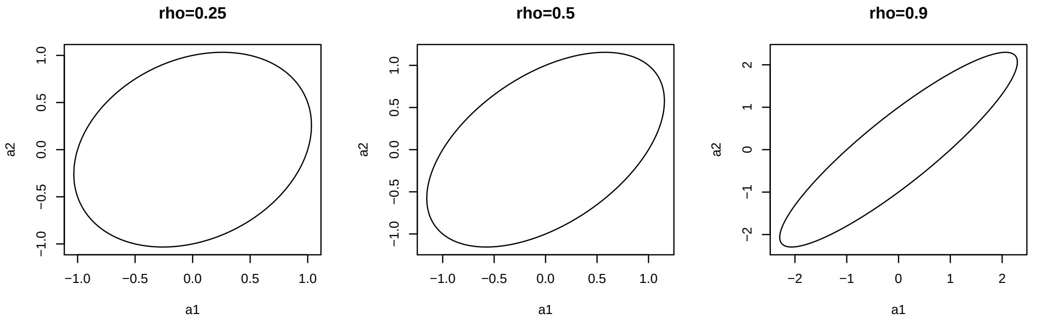

Figure 6.2: Plots of 2-dimensional ellipsoids for $\rho = 0.25;0.5;0.9$. Notice: each plot is differently scaled!

### 6.2.3 Theory and reality

The analytical results obtained above should be verified by checking if actual Hotelling's test outcomes correspond to the analytical results of real data. In this subsection we compared the behavior of squared Mahalanobis distance $T^{*2}$ with Hotelling's statistic $T^2$. For large $n_x$ and $n_y$ we assumed that $T^{*2} \approx T^2/c$, where $c = \frac{n_x n_y}{n_x+n_y}$. Constant $c$ changes as $n_x$ and $n_y$ change. It is reasonable to divide Hotelling's statistic $T^2$ by $c$ instead of multiplying $T^{*2}$ by $c$ in order to be able to compare how $T^2$ and $T^{*2}$ differ for various $n_x$ and $n_y$.

In order to compare the actual results with the analytical ones, we performed the following simulations. All data were simulated from $m$-dimensional normal distributions. We set the dimension $m$ to be 10, 15 and $m = 25$. All simulations were performed for three different values of the correlation coefficient $\rho$: $\rho = 0.1$, $\rho = 0.5$ and $\rho = 0.9$. In order to compare the behavior of Hotelling's test for various sizes of samples we took three choices of $n_x$ and $n_y$: $n_x = n_y = m, n_x = n_y = 1.4m$ and $n_x = n_y = 2.4m$. The number of false marginal distributions $k$ varied from one to $m$. The shift value for each of the different marginal distributions was set to one. The squared Mahalanobis distance was calculated according to (6.3). Hotelling's statistic is estimated from 1000 simulations for each case (as the mean of $T^2/c$ obtained from the simulations).

Plots of our simulated cases are shown in figure 6.3. We can see that for all simulated situations, the shapes of the squared Mahalanobis distance and Hotelling's statistics are similar. The only difference is in the heights of these curves. For small $n_x$ and $n_y$ statistic $T^2$ has higher values than for large $n_x$ and $n_y$. The reason for that stems from the inaccurate estimates of the expected values and the covariance matrix. However, we observe that with the increase of $n_x$ and $n_y$, statistic $T^2/c$ goes to $T^{*2}$ relatively fast. Therefore, the behavior of Hotelling's test for real data is expected to be very similar to the behavior of squared Mahalanobis distance $T^{*2}$.

In the previous section we saw that for the two-dimensional case the plotted shifts with equal values of the power of theoretical Hotelling's test form elliptic curves. Hotelling's statistics $T^2$ are random variables. Therefore, we can only estimate if their expected values form elliptic curves when they are plotted. To check this we performed the following simulation. Instead of calculating the shifts for which Hotelling's test has equal powers, we took some points with coordinates $(a_1, a_2)$ from the elliptic curves observed for squared Mahalanobis distance. For each such point, we did 1000 simulations and calculated Hotelling's statistic. We estimated the expected value $\mathrm{E}\, T^2/c$ as the mean for these 1000 repetitions. We divided Hotelling's statistics by $c$ for better understanding how fast these statistics go to $T^{*2}$. We did this simulation for the values of the correlation coefficient $\rho = 0.3$ and $\rho = 0.9$ and as the number of observations in each sample we took $n_x = n_y = 5$, $n_x = n_y = 10$ and $n_x = n_y = 20$. Results of our simulation are given in table 6.1. We observe that estimated mean values of $T^2/c$ are not very different, they go to $T^{*2}$ and their variance decreases with increasing number of observations. Clearly, these points form elliptic curves. Hence, we can claim that the real Hotelling's test behaves very similar to the theoretical one and the theory derived

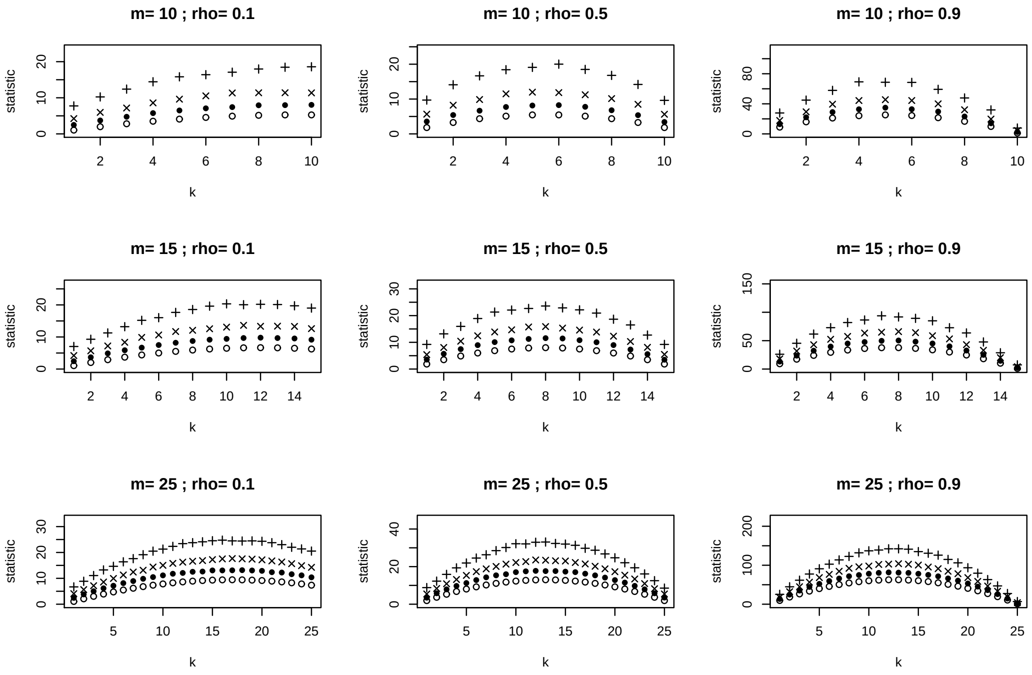

Figure 6.3: Comparisons of squared Mahalanobis distance $T^{*2}$ and real Hotelling's statistic $T^2/c$ for the dimension $m = 10$ 15, 25 (from the top to the bottom); for correlation coefficient $\rho = 0.1, 0.5, 0.9$ (from the left to the right) and number of observations in each sample $n_x = n_y = m$ (denoted by '+'), $n_x = n_y = 1.4m$ (denoted by 'x') and $n_x = n_y = 2.4m$ (denoted by '●'). Squared Mahalanobis distance $T^{*2}$ is denoted by '○'. The number of different marginal distributions $k$ is set from one to $m$. Notice: each plot is differently scaled!

for the theoretical test holds for the real Hotelling's test as well.

## 6.3 Comparison of tests for gene sets

Although there exist some papers (e.g. *Ackermann and Strimmer* (2009) and *Glazko and Emmert-Streib* (2009)) which compare tests for gene sets we performed our own simulation study. We considered two-sample problem and as gene sets we took independent samples of random vectors having $m$-dimensional normal distribution $N_m(0, \Sigma)$ and $N_m(\mu_2, \Sigma)$, with sample sizes $n_1$ and $n_2$, respectively. For simplicity we assumed that all elements on the main diagonal of the covariance matrix $\Sigma$ for both samples were

| $T^{*2} = 1.0989$ | | $\rho = 0.3$ | | | $T^{*2} = 5.2632$ | | $\rho = 0.9$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $n_s = 5$ | $n_s = 10$ | $n_s = 20$ | $a_1$ | $a_2$ | $n_s = 5$ | $n_s = 10$ | $n_s = 20$ |
| -0.84 | 0.35 | 3.12 | 1.74 | 1.35 | -1.83 | -1.05 | 9.58 | 6.72 | 5.96 |
| -0.63 | 0.61 | 3.03 | 1.81 | 1.42 | -1.38 | -0.44 | 9.55 | 6.51 | 5.96 |
| -0.42 | 0.79 | 3.04 | 1.82 | 1.39 | -0.92 | 0.09 | 9.55 | 6.65 | 5.99 |
| -0.21 | 0.92 | 3.00 | 1.75 | 1.42 | -0.46 | 0.57 | 9.62 | 6.93 | 5.98 |
| 0.00 | 1.00 | 3.03 | 1.72 | 1.42 | 0.00 | 1.00 | 9.10 | 6.99 | 5.83 |
| 0.21 | 1.04 | 3.04 | 1.74 | 1.36 | 0.46 | 1.39 | 9.74 | 6.78 | 5.99 |
| 0.42 | 1.04 | 3.01 | 1.87 | 1.39 | 0.92 | 1.74 | 10.11 | 6.75 | 5.86 |
| 0.63 | 0.99 | 3.00 | 1.79 | 1.40 | 1.38 | 2.04 | 9.36 | 6.87 | 5.85 |
| 0.84 | 0.85 | 3.32 | 1.81 | 1.41 | 1.83 | 2.25 | 10.21 | 6.87 | 5.96 |
| 1.05 | 0.35 | 3.35 | 1.85 | 1.36 | 2.29 | 2.09 | 9.94 | 6.85 | 5.97 |
| var: | | 0.0176 | 0.0025 | 0.0007 | var: | | 0.1133 | 0.0202 | 0.0039 |

Table 6.1: Results of simulations of two-dimensional adjusted Hotelling's statistics $T^2/c$ with $n_s = n_x = n_y$ observations for each sample and correlation coefficient $\rho$. $T^{*2}$ stands for squared Mahalanobis distance and $(a_1, a_2)$ is difference between expected values $\mu_x - \mu_y$ of these samples. On bottom line is the estimate of variance of each column.

equal to 1 and all other elements were equal to $\rho > 0$, i.e.

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \dots & \rho & 1 & \rho \\ \rho & \dots & \dots & \rho & 1 \end{pmatrix}.$$

Further on, we assumed that $\mu_2$ has the first $k$ elements equal to $\mu$ and the others equal to 0, i.e.

$$\mu_2 = (\underbrace{1, \dots, 1}_{k}, \underbrace{0, \dots, 0}_{m-k})^T,$$

where $k$ is the number of differentially expressed genes. We set the number of observations in each group to be $n = 20, 40, 100$. The number of genes in gene sets was $m = 20, 50, 100$ with restriction $2n > m$ and the number of differentially expressed genes $k$ between these two groups was set to be integer part of $1, m/4, m/2, m/4, m$. We took correlation coefficient $\rho$ equal to $0.1, 0.5, 0.9$ and as the difference parameter $\mu$ we took sequence of eleven equidistant values begin from 0 (different for each $n$ and $\rho$). The $p$-values for considered tests were based on 1000 permutations and each simulation was repeated 1000 times. The power of tests was estimated as an average of rejections from these 1000 repetitions.

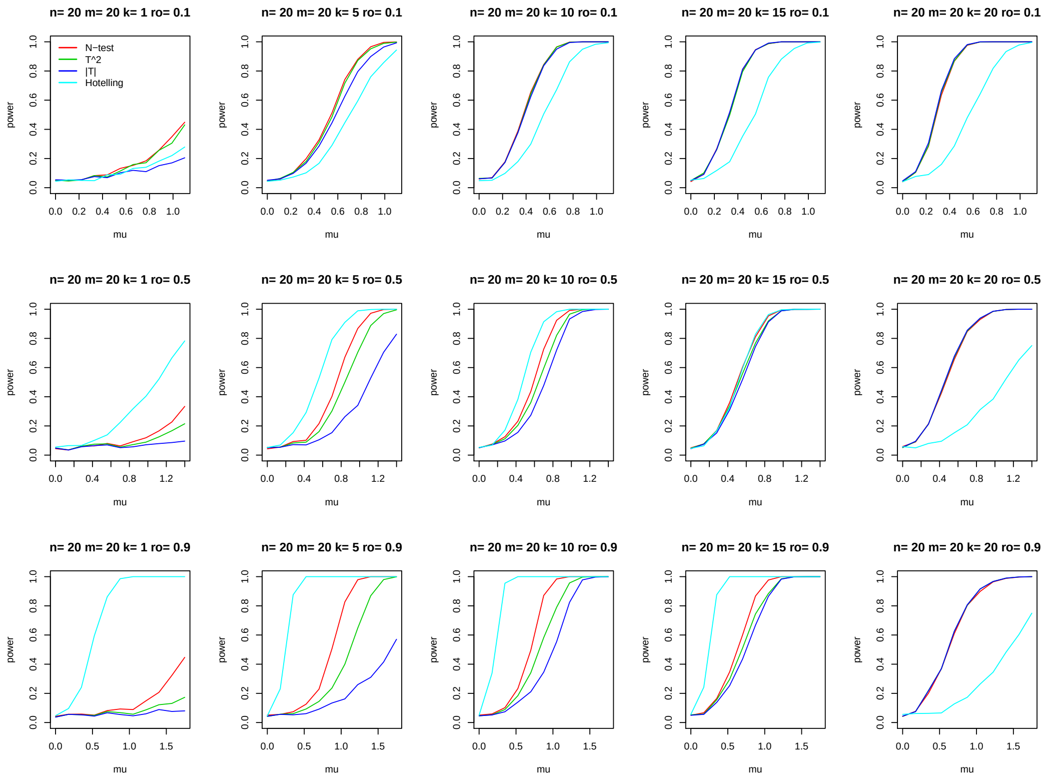

Figure 6.4: Comparison of test for gene sets for $n = 20$, $\rho = 0.1, 0.5, 0.9$ and $m = 1, 5, 10, 15, 20$.

Results of this simulation for $n = 20$ and $m = 20$ are in figure 6.4. The whole results can be found in the supplement of this work. We can see that with increasing number of differentially expressed genes $k$ the power of $N$-test and both tests based on $t$-statistics increases. On the other hand, their power decreases with increasing $\rho$. $N$-test has the best power between these three tests, but for $k = m$, the power of these three tests is almost equivalent. The Hotelling's test behaves different way. For small $\rho$ it has the poorest power among all four tests. However, with higher $\rho$ its power increases and this test is the most powerful except for large $k$, where its power decreases too much. This behavior of Hotelling's test was discussed in the previous section. However, for gene expression data, there are expected high correlations between genes and just small number of differently expressed genes. Therefore, Hotelling's test seems to be the best for such data.

# Chapter 7

# Dependence vs Correlativity

It was shown that normalization of gene expressions data makes these data almost uncorrelated. However, if correlation coefficient between two random variables is equal to zero, it does not generally mean that these variables are independent. It is only true for random variables having normal distribution. Therefore, in this chapter, we derived test for testing independence of two random samples based on empirical characteristic functions. Moreover, we study the power properties of this test.

For gene expressions there were derived two types of dependence between genes: type A dependence (see *Klebanov et al.* (2006)) and hidden regulator dependence (see *Lim et al.* (2010)). The dependence structure of genes could be very important in practice. Therefore, at the end of this chapter we show that HYPERDIP and TEL have much more genes with type A dependence than genes with hidden regulator dependence.

## 7.1 Test statistic

Consider that we have two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from distributions (say $X$ and $Y$, respectively) with the characteristic functions $f_X(s) = Ee^{isX}$ and $f_Y(s) = Ee^{isY}$, respectively. Moreover, consider that $f_{XY}(s,t) = Ee^{isX+itY}$ is characteristic function of random vector (X,Y). Then these two distributions are independent if and only if $f_{XY}(s,t) = f_X(s) f_Y(t)$. We can employ this knowledge and use test statistic based on difference of $f_{XY}(s,t) - f_X(s) f_Y(t)$. Because these characteristic functions are unknown, we estimate them by their empirical counterparts. It leads us to measure the amount of dependence between $X$ and $Y$ by statistic

$$\phi(X,Y) = \max_{-c \le s,t \le c} \left| \frac{1}{n} \sum_{i=1}^{n} e^{isX_i + itY_i} - \frac{1}{n^2} \sum_{i=1}^{n} e^{isX_i} \sum_{i=1}^{n} e^{itY_i} \right|, \tag{7.1}$$

where $c > 0$ is some border constant. We will call this statistics $\phi$-statistic (test based on this statistic we will call the $\phi$-test).

Another alternative to the $\phi$-test could be based on statistic

$$\phi^*(X,Y) = \max_{-c \leq s,t \leq c} \left| \frac{\frac{1}{n}\sum_{i=1}^n e^{isX_i+itY_i} - \frac{1}{n^2}\sum_{i=1}^n e^{isX_i}\sum_{i=1}^n e^{itY_i}}{st} \right|. \tag{7.2}$$

The only difference between $\phi$-statistic and $\phi^*$-statistic is in the denominator of absolute value. For $\phi$ it is equal to one, for $\phi^*$ it is equal to $st$. We consider $-c \leq s, t \leq c$ because empirical characteristic functions are periodical.

Characteristic functions depend on mean and variance of random samples. Hence, $\phi$-statistic and $\phi^*$-statistic are affected by variance and mean of both samples. Therefore, the samples $X$ and $Y$ should be standardized to have mean value equal to zero and variance equal to one. Then we should work with standardized samples instead of original ones. Such standardization could be done by subtracting the sample mean and then dividing by sample standard deviation.

Now we try to find out which value is optimal for border constant $c$ and if $\phi$ is better than $\phi^*$ or not. The exact distribution of $\phi$-statistic and $\phi^*$-statistic is not known. Therefore, we performed simple simulation in order to estimate 95% quantile of $\phi$-statistic and $\phi^*$-statistic for different border constant $c$ and different number of observations $n$ based on 10000 repetitions. Instead of computing maximum in (7.1) numerically we made a square lattice with mesh size 0.01x0.01 for $c = 0.1, 0.3, 0.5, 1$ and we calculated maximum of $\phi$ on it. For $\phi^*$ we took square lattice with mesh size 0.01x0.01 for $c = 0.1$ and $c = 0.3$ and moreover for $c^* = 0.1$ we took square lattice with mesh size 0.0025x0.0025 as well.

Firstly, we simulated random samples from two-dimensional normal distribution with mean equal to zero and variance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

We set the number of observations $n = 10, 35, 60, 79$ and correlation coefficient $\rho = 0, 0.3, 0.6, 0.9$. For each setting we standardized the samples and compute $\phi$-statistic and $\phi^*$-statistic. We repeated all setting 10000 times and estimated power of $\phi$-test and $\phi^*$-test. Results of this simulation are in table 7.1. We can see that the power is similar for all cases except for $\phi$-test with $c = 1$ where the power is weaker (and for larger $c$, the power would be much weaker). Moreover with increasing correlation coefficient the power of $\phi$-test and $\phi^*$-test increases.

## Csorgo's test

In paper *Csörgö* (1985), there was derived test (we call it Csorgo's test) for testing independence of two samples using characteristic functions. To find out whether $\phi$-test

|  | $n = 10$ |  |  | $n = 35$ |  |  |
|---|---|---|---|---|---|---|
|  | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ |
| $\phi$-test |  |  |  |  |  |  |
| $c = 0.1$ | 0.136 | 0.497 | 0.985 | 0.434 | 0.979 | 1.000 |
| $c = 0.3$ | 0.129 | 0.477 | 0.983 | 0.424 | 0.976 | 1.000 |
| $c = 0.5$ | 0.130 | 0.492 | 0.985 | 0.434 | 0.978 | 1.000 |
| $c = 1$ | 0.117 | 0.418 | 0.969 | 0.303 | 0.925 | 1.000 |
| $\phi^*$-test |  |  |  |  |  |  |
| $c = 0.1$ | 0.137 | 0.494 | 0.983 | 0.430 | 0.980 | 1.000 |
| $c = 0.3$ | 0.132 | 0.487 | 0.985 | 0.432 | 0.980 | 1.000 |
| $c = 0.1^*$ | 0.135 | 0.495 | 0.984 | 0.425 | 0.979 | 1.000 |
|  | $n = 60$ |  |  | $n = 79$ |  |  |
|  | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ |
| $\phi$-test |  |  |  |  |  |  |
| $c = 0.1$ | 0.658 | 1.000 | 1.000 | 0.776 | 1.000 | 1.000 |
| $c = 0.3$ | 0.658 | 1.000 | 1.000 | 0.791 | 1.000 | 1.000 |
| $c = 0.5$ | 0.654 | 1.000 | 1.000 | 0.762 | 1.000 | 1.000 |
| $c = 1$ | 0.481 | 0.994 | 1.000 | 0.585 | 0.999 | 1.000 |
| $\phi^*$-test |  |  |  |  |  |  |
| $c = 0.1$ | 0.656 | 0.999 | 1.000 | 0.782 | 1.000 | 1.000 |
| $c = 0.3$ | 0.661 | 1.000 | 1.000 | 0.780 | 1.000 | 1.000 |
| $c = 0.1^*$ | 0.663 | 1.000 | 1.000 | 0.780 | 1.000 | 1.000 |

Table 7.1: Power of $\phi$-test and $\phi^*$-test for normal distribution calculated on square lattice with mesh size 0.01x0.01 (last case for $\phi^*$ denoted by $c^*$ is with mesh size 0.0025x0.0025).

| Csorgo's test | $\rho = 0$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ |
|---|---|---|---|---|
| $n = 10$ | 0.1267 | 0.1222 | 0.2093 | 0.3789 |
| $n = 35$ | 0.1191 | 0.2195 | 0.5406 | 0.7129 |
| $n = 60$ | 0.0999 | 0.3108 | 0.7363 | 0.8911 |
| $n = 79$ | 0.0652 | 0.3890 | 0.8434 | 0.976 |

Table 7.2: Power of Csorgo's test for two dimensional normal distribution with correlation coefficient $\rho = 0, 0.3, 0.6, 0.9$.

has good power we try to compare this test with Csorgo's test. Hence we did similar simulation study for Csorgo's test as we did for $\phi$-test. We set the number of observations $n = 10, 35, 60, 79$ and correlation coefficient $\rho = 0, 0.3, 0.6, 0.9$. For each setting we standardized the samples and estimate the power of Csorgo's test based on 10000 repetitions. Results of these simulations are in table 7.2. We can see that for small number of observations this test does not hold significance level $\alpha$. Moreover it has poor power in comparison with $\phi$-test. Therefore, we can claim that $\phi$-test has good power properties.

## 7.2 Non-normal distribution

Now we consider Laplace distribution. We simulated similar cases as for normal distribution. For considering dependent random samples from Laplace distribution we created $n$x2 matrix from these independent samples of sizes $n$ (each row is different sample) and multiplied it by matrix $\Sigma$ to create dependent samples. The power of $\phi$-test and $\phi^*$-test for Laplace distribution is in table 7.3. We can see that this power for independent random variables is about 5%. It means that these tests hold nominal level for Laplace distribution too. Moreover, power for dependent variables having Laplace distribution is similar to power for normal distributed random variables. In addition, power for all settings is similar except for $\phi$-test with $c = 1$ where the power is weaker.

From Bernstein theorem (see *Kagan et al.* (1973)) we know that random variables $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$ are uncorrelated and they are independent if and only if $X_1$ and $X_2$ have normal distribution. Therefore, for finding if $\phi$-value is related with correlation coefficient or with independence, we performed simple simulation. For $n = 10, 35, 60, 79$ we took independent samples $N_1$, $N_2$ from $N(0, 1)$, independent samples $L_1$, $L_2$ having central Laplace distribution with variance equal to one and independent samples $B_1$, $B_2$ from Bernoulli distribution with parameter $\epsilon = 0, 0.1, \ldots, 1$. We took samples $X_1 = (1 - B_1)N_1 + B_1L_1$ and $X_2 = (1 - B_2)N_2 + B_2L_2$. For $\epsilon = 0$, samples $X_1$ and $X_2$ were independent having normal distribution and for $\epsilon = 1$ we had independent samples from Laplace distribution. We calculated samples $Y_1 = X_1 + X_2$

|  | $n = 10$ |  |  |  | $n = 35$ |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | $\rho = 0$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ | $\rho = 0$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ |
| $\phi$-test |  |  |  |  |  |  |  |  |
| $c = 0.1$ | 0.054 | 0.145 | 0.546 | 0.990 | 0.049 | 0.441 | 0.981 | 1.000 |
| $c = 0.3$ | 0.050 | 0.135 | 0.517 | 0.988 | 0.045 | 0.437 | 0.984 | 1.000 |
| $c = 0.5$ | 0.053 | 0.140 | 0.528 | 0.987 | 0.044 | 0.415 | 0.979 | 1.000 |
| $c = 1$ | 0.039 | 0.099 | 0.410 | 0.975 | 0.024 | 0.208 | 0.893 | 1.000 |
| $\phi^*$-test |  |  |  |  |  |  |  |  |
| $c = 0.1$ | 0.055 | 0.136 | 0.526 | 0.988 | 0.051 | 0.444 | 0.979 | 1.000 |
| $c = 0.3$ | 0.053 | 0.141 | 0.526 | 0.987 | 0.059 | 0.446 | 0.981 | 1.000 |
| $c = 0.1^*$ | 0.056 | 0.148 | 0.537 | 0.985 | 0.052 | 0.435 | 0.981 | 1.000 |
|  | $n = 60$ |  |  |  | $n = 79$ |  |  |  |
|  | $\rho = 0$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ | $\rho = 0$ | $\rho = 0.3$ | $\rho = 0.6$ | $\rho = 0.9$ |
| $\phi$-test |  |  |  |  |  |  |  |  |
| $c = 0.1$ | 0.047 | 0.651 | 0.999 | 1.000 | 0.053 | 0.788 | 1.000 | 1.000 |
| $c = 0.3$ | 0.047 | 0.662 | 0.999 | 1.000 | 0.049 | 0.794 | 1.000 | 1.000 |
| $c = 0.5$ | 0.042 | 0.635 | 0.999 | 1.000 | 0.036 | 0.747 | 1.000 | 1.000 |
| $c = 1$ | 0.025 | 0.342 | 0.992 | 1.000 | 0.021 | 0.427 | 0.999 | 1.000 |
| $\phi^*$-test |  |  |  |  |  |  |  |  |
| $c = 0.1$ | 0.049 | 0.676 | 0.999 | 1.000 | 0.054 | 0.793 | 1.000 | 1.000 |
| $c = 0.3$ | 0.048 | 0.672 | 1.000 | 1.000 | 0.051 | 0.789 | 1.000 | 1.000 |
| $c = 0.1^*$ | 0.056 | 0.699 | 1.000 | 1.000 | 0.050 | 0.783 | 1.000 | 1.000 |

Table 7.3: The power of $\phi$-test for Laplace distribution calculated on square lattice with mesh size 0.01x0.01 (last case for $\phi^*$ denoted by $c^*$ is with mesh size 0.0025x0.0025).

and $Y_2 = X_1 - X_2$. In figure 7.1, there is the power of $\phi$-test for $c = 0.5$ of samples $Y_1$ and $Y_2$ for different $\epsilon$. In table 7.4, there are results for $\epsilon = 0$ and $\epsilon = 1$ for all setting of $\phi$-test and $\phi^*$-test. We can see, that for $\epsilon = 0$ (for independent samples) the power is about 0.05 for both tests. With increasing $\epsilon$ (and with increasing dependence as well) the power of our tests increases. It means that $\phi$-test and $\phi^*$-test, respectively are related with dependence and not only with correlation coefficient.

For Bernstein case, the power increases as $c$ increases. The power of $\phi$-test for $c = 0.1, 0.3, 0.5$ and for $\phi^*$-test (for all settings of $c$) was similar. From these settings for Bernstein case, the $\phi$-test with $c = 0.5$ has higher power. Therefore, in the rest of this chapter we will use the $\phi$-test with border constant $c = 0.5$.

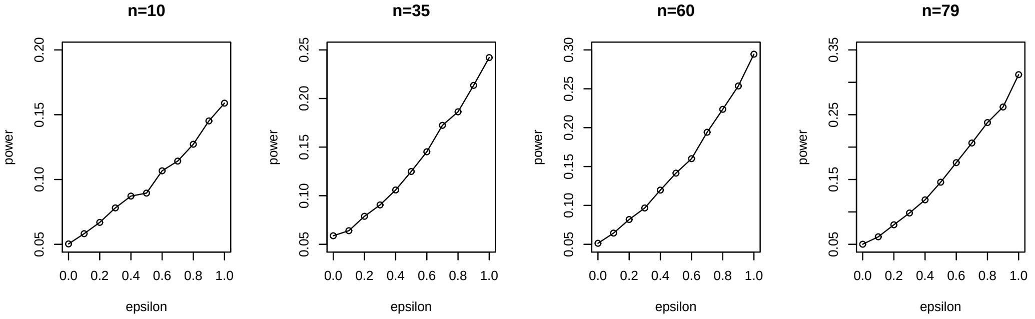

Figure 7.1: Plots of power for different $\epsilon$ for $c = 0.5$.

## 7.3 HYPERDIP and TEL data

Let us look at HYPERDIP and TEL data. In both data sets, we ordered genes in decreasing order of their estimated variance. Therefore, the gene with the highest estimated variance was the first and the gene with the smallest variance was the last. In figure 7.2, we can see estimates of variance according to their order. The variance decreases quickly only at the beginning and at the end and there are most of genes with similar variance.

Firstly, remember what type A dependence is. Let $X$ and $Y$ be gene $\log_2$ expression levels for gene $g_x$ and $g_y$, respectively. We say that pair $(g_x, g_y)$ is type A dependent if $X$ and $Y$ satisfy the condition $Y = X + Z$, where $Z$ is a random variable stochastically independent on $X$.

It is too time consumable to compute $\phi$-test for all pairs of genes. Therefore, we considered three parts of data. We took first 1000 genes with the highest variance, 1000 genes with the smallest variance and 1000 genes from the middle with indexes from 3001 to

| | normal | | | | laplace | | | |
|---|---|---|---|---|---|---|---|---|
| | $n = 10$ | $n = 35$ | $n = 60$ | $n = 79$ | $n = 10$ | $n = 35$ | $n = 60$ | $n = 79$ |
| $\phi$-test | | | | | | | | |
| $c = 0.1$ | 0.054 | 0.053 | 0.046 | 0.051 | 0.157 | 0.196 | 0.203 | 0.211 |
| $c = 0.3$ | 0.044 | 0.052 | 0.050 | 0.054 | 0.157 | 0.203 | 0.223 | 0.248 |
| $c = 0.5$ | 0.050 | 0.059 | 0.051 | 0.050 | 0.159 | 0.242 | 0.294 | 0.312 |
| $c = 1$ | 0.045 | 0.052 | 0.054 | 0.047 | 0.143 | 0.297 | 0.436 | 0.520 |
| $\phi^*$-test | | | | | | | | |
| $c = 0.1$ | 0.050 | 0.048 | 0.048 | 0.054 | 0.158 | 0.200 | 0.198 | 0.209 |
| $c = 0.3$ | 0.053 | 0.057 | 0.050 | 0.051 | 0.151 | 0.217 | 0.210 | 0.222 |
| $c = 0.1^*$ | 0.048 | 0.050 | 0.054 | 0.050 | 0.162 | 0.220 | 0.262 | 0.269 |

Table 7.4: Power of $\phi$-test for Berstein theorem (uncorrelated samples) and $\epsilon = 0$ (normal distribution - independent samples) and $\epsilon = 1$ (Laplace distribution - dependent samples) calculated on square lattice with mesh size 0.01x0.01 (last case for $\phi^*$ denoted by $c^*$ is with mesh size 0.0025x0.0025).

4000. Each group (called upper, middle and bottom) was investigated separately. So we created $3 \,\mathrm{x}\, 2 = 6$ data sets. We not only considered $\phi$-test for genes but $\phi$-test for $G_i$ and $G_j - G_i$, $j < i$, where $G_i$ denotes $i$-th gene. We call the second case the type A situation because if genes $G_i$ and $G_j$ are type A dependent, then $G_i$ and $G_j - G_i$ are independent. Therefore, $\phi(G_i, G_j - G_i)$ should take smaller values than $\phi(G_i, G_j)$. Because $\phi$-test is not variance invariant, we standardized each gene and each difference of genes so that they had zero mean and variance equal to one. The border constant $c$ was set to be 0.5. In figure 7.3, we can see histograms of $\phi$-values (499500 pairs on each plot) for all 12 situations. We can see that for type A situation $\phi$-values take lower values than for casual pairs of genes. These histograms are overlapped just for upper genes. However, for type A situation, there are 77.3% (for HYPERDIP) and 99.7% (for TEL) of $\phi$-values lower than 1%-quantile of $\phi$-value for casual pairs of genes. Therefore, type A pairs are much more independent than casual pairs.

Now consider mix of three groups of genes divided according their variance. We created new six data sets (three for HYPERDIP and three for TEL data). In order to create these 3 groups we created three subgroups created from 500 genes with highest variance, 500 genes indexed from 3251 to 3750 in decreasing order of variance and 500 genes with smallest variance, respectively. Three considered groups were created by taking two of three subgroups together. $\phi$-values in each subgroup had been computed in previous situations, therefore we considered only $\phi$-test for genes from different subgroups. Therefore, we computed $500 \,\mathrm{x}\, 500 = 250000$ $\phi$-values for each data set. In figure 7.4, there are histograms of $\phi$-values for each data set. Again, we can see that $\phi$-values for casual pairs are bigger than for type A situations. Non-ovelapping his-

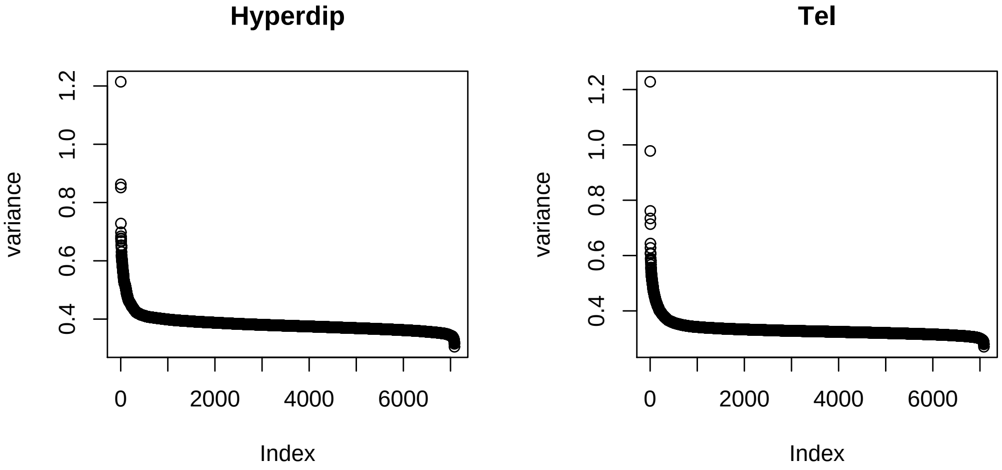

Figure 7.2: Plots of estimates of variance in decreasing order.

tograms are just for middle-bottom situation. Therefore, we are interested in computing proportions of $\phi$-values for type A situations lower than 1%-quantile of $\phi$ for casual pairs. Values of these proportions are in table 7.5. Again we can say that type A pairs are much more independent than pairs of genes in casual case. Moreover, in table 7.6 there are proportions of hypotheses we reject according to $\phi$-test on 5% nominal level for casual pairs and for type A situation. We can see that almost all hypotheses for casual cases are rejected and there are many hypotheses which failed to reject for type A situation.

| | upper-mid | upper-bottom |
|---|---|---|
| HYPERDIP | 0.897 | 0.944 |
| TEL | 0.998 | 0.988 |

Table 7.5: Proportions of $\phi$-values for type A situation lower than 1%-quantile of $\phi$-values for pairs of genes for upper-mid, upper-bottom and mid-bottom data sets.

## 7.4 Hidden regulator dependence

*Lim et al.* (2010) considered another type of dependence between genes called hidden regulator dependence (HRD). They considered two genes (say X and Y) being HRD if

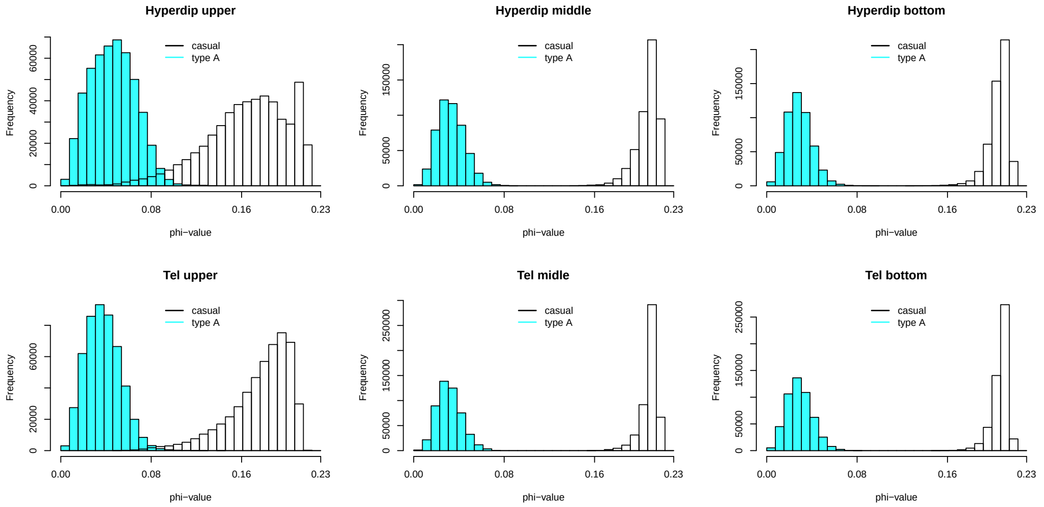

Figure 7.3: Histograms of $\phi$-values for casual pairs of genes $G_i$ and $G_j$; $i < j$ (black color) and for type A situation (aqua color) for HYPERDIP and TEL data. We consider 1000 genes with highest variance (upper), 1000 genes with indexes 3001 to 4000 in decreasing order of variance (middle) and 1000 genes with smallest variance (bottom).

there exists random variable $a$ such that for $\log_2$-expression $x$ and $y$ holds $x = a + \epsilon_x$ and $y = a + \epsilon_y$, where $\epsilon_x$ and $\epsilon_y$ are i.i.d random variables independent on $a$. They numerically demonstrated that HRD is easily mistaken for type A dependence. Consider two genes, say $G_i$ and $G_j$. If these genes are HRD then it does not matter if we calculate $\phi$-test for $G_i$ and $G_j - G_i$ or $G_j$ and $G_j - G_i$ respectively. But if these genes are type A dependent then it is not true and $\phi$-values for $G_i$ and $G_j - G_i$ or $G_j$ and $G_j - G_i$ are expected different because one pair is independent and the other not. Therefore, we calculate $\phi$-test for both pairs. By sorting genes in decreasing order of their estimated variance we expect that for $j < i$ the pair $G_i$ and $G_j - G_i$ is independent and the pair $G_j$ and $G_j - G_i$ is dependent. Therefore, we only take pairs of genes for which we consider type A dependence ($\phi$-test does not reject independence of $G_i$ and $G_j - G_i$, $j < i$ at nominal level 5%). For these pairs we calculate $\phi$-test for $G_j$ and $G_j - G_i$. In table 7.7, there are proportions of such pairs for which $\phi$-test is rejected. We can see that a lot of pairs are rejected according to $\phi$-test (for mixed groups this proportion is almost one). Therefore, we can see that type A dependence exists for genes with very different variances not as HRD which can exist only for genes with similar estimate of variances.

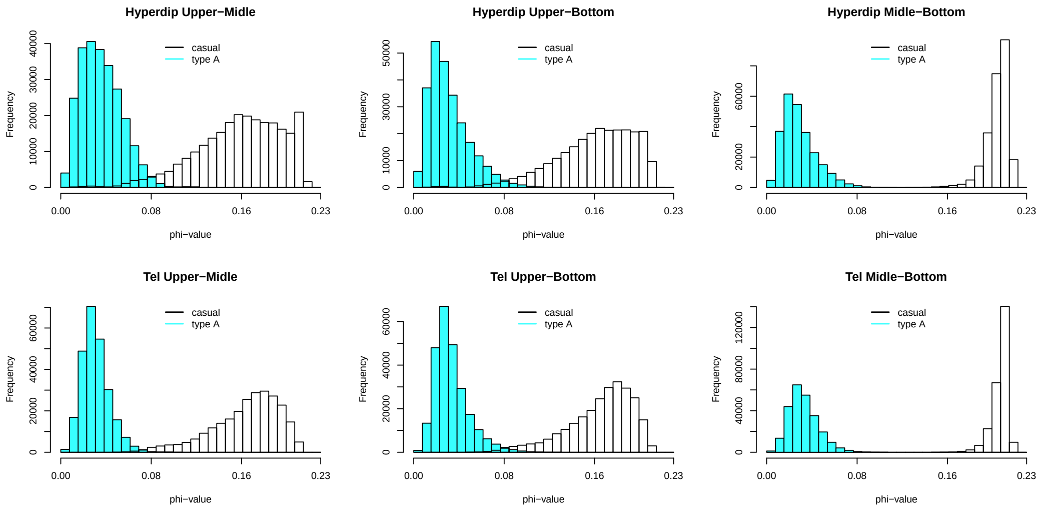

Figure 7.4: Histograms of $\phi$-values for casual pairs of genes $G_i$ and $G_j$; $i < j$ (black color) and for type A situation (aqua color) for HYPERDIP and TEL data. We consider plots for genes from upper-mid data sets, upper-bottom data sets and mid-bottom data sets.

| | HYPERDIP | | | TEL | | |
|---|---|---|---|---|---|---|
| | Upper | Mid | Bottom | Upper | Mid | Bottom |
| normal | 0.9957 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| type A | 0.5140 | 0.1596 | 0.0777 | 0.2553 | 0.0798 | 0.0579 |
| | U-B | U-M | M-B | U-B | U-M | M-B |
| normal | 0.9958 | 0.9954 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| type A | 0.2006 | 0.2935 | 0.1430 | 0.1526 | 0.0952 | 0.1272 |

Table 7.6: Proportion of rejected hypotheses for HYPERDIP a TEL data for normal pairs and type A pairs.

| HYPERDIP | | | TEL | | |
|---|---|---|---|---|---|
| Upper | Mid | Bottom | Upper | Mid | Bottom |
| 0.761 | 0.179 | 0.489 | 0.885 | 0.068 | 0.400 |
| U-B | U-M | M-B | U-B | U-M | M-B |
| 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 |

Table 7.7: Proportion of type A genes for which $G_j$ and $G_j - G_i, j < i$ are considered dependent.

# Chapter 8

# Conclusion

One of the goals of statistician working with microarray data is to find differentially expressed genes between two groups of observations (e.g. treatment versus control conditions, two stages of some illness). In this work, we tried to solve some problems that make finding differentially expressed genes difficult.

To find differentially expressed genes one has to use two-sample test for testing equality of means or equality of distributions between two samples of genes. Gene expressions after $\log_2$ transformation are considered to have approximately normal distribution. Therefore, in many cases $t$-test can be useful. The problem set in when we cannot guarantee normality or there can be change not only in mean but in variance of distributions as well. Then we could use some non-parametric test such as two-sample Kolmogorov-Smirnov test. In this work, we showed that this test could be biased in case that there is different number of observations in each sample of genes. Moreover, this test lacks of power. Due to the luck of power of two-sample Kolmogorov-Smirnov test we propose to use $N$-test. This test is distribution free, we found out that it has good power properties with comparison to $t$-test and it is much better than $t$-test in case that there is difference in variance of two samples. Unfortunately, we do not know the distribution of $N$-test. The $p$-values of this test have to be computed by permutations what makes this test too time consumable. Therefore, we should still prefer $t$-test when we need to compute $p$-values of hypotheses quickly.

One of the problems of microarray data is that we usually have a large number of genes. Hence, we have to compare thousands of genes simultaneously. In other words, we have to test thousands of hypotheses simultaneously. Therefore, some multiple testing procedure has to take place here. Bonferroni procedure is well-known procedure but it is considered to be too conservative and therefore to have weak power. In *Gordon* (2007), they proved that this procedure is unimprovable in the class of monotone step-up multiple testing procedures controlling $FWER$ for each dependence structure of $p$-values. Holm procedure is step-down improvement of Bonferroni procedure. This procedure dominates all monotone step-down procedures controlling $FWER$ and is unimprovable in the class of monotone multiple testing procedures controlling $FWER$. Our simulations showed that there is only small difference in estimates of $FWER$ of these two procedures and the lines of power are almost overlapped. Therefore, the conservativeness of Bonferroni procedure comes from principle of multiple testing and not from simplicity of this procedure. The improvement in power can be achieved by using procedures controlling $FDR$. Although these procedures find more differentially expressed genes, they produce more false discoveries as well. Benjamini-Yekutiely procedure is too conservative in controlling $FDR$. Therefore, empirical Bayes approach can help here. On the other side, it produces too many false positives. Hence, we have to consider if we want to control $FWER$ (find any false positives with small probability) and use Bonferroni or Holm procedure or we want to control $FDR$ (find more differentially expressed genes but produce a large number of false positives as well) and use empirical Bayes approach.

Another problem of gene expression data is that genes are highly correlated between themselves. Hence, $p$-values of hypotheses about genes are dependent. It influences their properties as we showed in chapter 3. There exist some normalizations that partially solve this problem. The most common normalization such as global normalization or quantile normalization makes normalized genes almost uncorrelated. On the other hand, they influence both differentially and non-differentially expressed genes. Therefore, they result in finding too many false positives. In other words, genes that seem to be differentially expressed after these normalizations do not need to be differentially expressed in original samples. In this work, we proposed some normalizations based on $\delta$-sequence of *Klebanov and Yakovlev* (2007). We showed that our proposal $D$ results in appropriate number of false positives (another normalizations do not). Moreover, after this normalization, tests discover more truly differentially expressed genes than if we use non-normalized data.

In some situations, it is better to work with genes sets instead of genes alone. Therefore, we test two-sample hypotheses about these sets and thus dealing with multidimensional hypotheses. The common way is to use Hotelling's test or tests based on $t$-statistic. These tests assume normality of samples. Moreover, Hotelling's test assumes equal covariance matrix for both samples. In this work, we showed that Hotelling's test has different behavior for dependent components of observations as another tests. Hotelling's test has good power only if there is small proportion of differentially expressed genes and it lacks the power if there is high proportion of differentially expressed genes in gene sets. Good alternative to this test is $N$-test, which is more powerful than tests based on $t$-statistic. On the other hand, $N$-test is too time consumable and it makes trouble to use this test, especially when we cannot afford long computation of $p$-values.

Finally, we know that normalizations make gene expression data almost uncorrelated. However, if data are uncorrelated it does not necessarily mean that these data are independent. Common tests usually test only uncorrelation and not independence of the data. Therefore, we proposed $\phi$-test to test independence of genes. This test helped us distinguish between type A dependence and hidden regulator dependence which can occur in gene expression data.

# Appendix A

# Useful inequalities

**Bonferroni inequality (known as Boole inequality as well):** Consider set of events $B_1, \ldots, B_n$. Then, Bonferroni inequality states that

$$P(\bigcup_{i=1}^{n} B_i) \leq \sum_{i=1}^{n} P(B_i). \tag{A.1}$$

**Markov's inequality:** If $X$ is any random variable and $a > 0$, then

$$P(|X| \geq a) \leq \frac{E|X|}{a}. \tag{A.2}$$

*p*-**value inequality:** The *p*-value satisfies the following inequalities with respect to true hypothesis $H$

$$P_H(p \leq \alpha) \leq \alpha. \tag{A.3}$$

# Appendix B

# Multivariate normal distribution with dependent components

For generating i.i.d random vectors $X_1, \ldots, X_n$, $X_j = (x_{1j}, \ldots, x_{mj})$, $j = 1, \ldots, n$ from $m$-dimensional normal distribution with zero mean and covariance matrix given by

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \ldots & \rho & 1 & \rho \\ \rho & \ldots & \ldots & \rho & 1 \end{pmatrix}$$

we use the following algorithm.

**Algorithm B.1.**

1. *Generate independent random variables $a_j$ and $y_{ij}$, $i = 1, \ldots, m$; $j = 1, \ldots, n$ from the standard normal distribution.*

2. *For a fixed $\rho$, define $x_{ij} = \sqrt{\rho}\, a_j + \sqrt{1-\rho}\, y_{ij}$, $i = 1, \ldots, m$; $j = 1, \ldots, n$.*

This algorithm produces random vectors with central normal distribution and moreover for each $i_1 \neq i_2$ all pairwise correlations $corr(x_{i_1 j}, x_{i_2 j}) = \rho$ and for each $i, k = 1, \ldots, m$ and $j \neq l$ we have $corr(x_{ij}, x_{k,l}) = 0$.

# Appendix C

# Supplement

In this work, we performed many simulations. Due to their complexity, all results cannot be inserted in this work. However, omitted results can be found on supplement CD which contains directories (they have the same name as the section they belong to)

- **1.6 HYPERDIP and TEL data** - in *HYPERDIP.txt* and *TEL.txt* files there are HYPERDIP and TEL data that are used across this doctoral thesis.

- **3.2 Some hypotheses are false** - in *Proportions of genes.pdf* and *Pairs of genes.pdf* files there are histograms of *p*-values of proportions of genes and pairs of genes according to different ordering.

- **4.2.1 Comparison of Bonferroni and Holm procedure** - in *Bonferroni-Holm.pdf* file there are complete results of simulation of section 4.2.1.

- **4.5.1 Comparison of empirical Bayes approaches** - in *EB comparison.pdf* file there are complete results of simulation of section 4.5.1.

  **4.6 Comparison of multiple testing procedures** - in *EB&MTP comparison.pdf* file there are complete results of simulation of section 4.6.

  **5.3 Comparison of normalizations** - in *Estimate of average power.pdf, Estimate of FDR.pdf, Estimate of FWER.pdf, Estimate of PFER.pdf* files there are tables of complete results of simulation of section 5.3.

- **6.3 Comparison of tests for gene sets** - in *Tests for gene sets.pdf* file there are complete results of simulation of section 6.3.

# Bibliography

Ackermann, M. and Strimmer, K.(2009), A general modular framework for gene set enrichment analysis, *BMC Bioinformatics*, 10, 47.

Barry, W.,T., Nobel, A., B. and Wright, F., A. (2008), A statistical framework for testing functional categories in microarray data, *The Annals of Applied Statistics*, 2 No.1, 286-315.

Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of Royal Statistical Society, Series B*, 57, 289-300.

Benjamini, Y. and Yekutieli, D. (2001), The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics*, 29, 1165-1188.

Bolstad, M., Irizarry, R., Strand, M. and Speed, T. (2003), A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, 19, 185-193.

Bonferroni, C., E. (1936), Teoria statistica delle classi e calcolo delle probabilia, *Publicazioni del R Instituto Superiore di Scienze Economiche e Commerciali di Firenze*, 3-62.

Bubeliny, P. (2008), Proportions of Gene Expressions in Microarray Data Analysis, *WDS'08 Proceedings of Contributed papers: Part I - Mathematics and Computer Sciences*, 106-111.

Bubeliny, P. (2011), Hotelling Test for Highly Correlated Data, *Acta Universitatis Carolinae. Mathematica et Physica*, 52 No.2, 67-75.

Bubeliny, P. (2013a), A Note on the Biasedness and Unbiasedness of Two-sample Kolmogorov-Smirnov Test, *Statistics & Risk Modeling*, 30 vol. 2, 181-188.

Bubeliny, P. (2013b), Using of Normalizations for Gene Expression Analysis, *Statistical Methods for Microarray Data Analysis, Humana press*, 73-83.

Chatfield, C. and Collins, A.,J. (1980), Introduction To Multivariate Analysis, *Chapman&Hall/CRC*.

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

Csörgö, S. (1985), Testing for independence by the empirical characteristic function, *Journal of Multivariate Analysis*, 16, 290-299.

Cui, X. and Churchill, G., A. (2003), Statistical tests for differential expression in cDNA microarray experiments, *Genome Biology*, 4, 4:201.

D'haeseleer, P., Liang, S. and Somogyi, R. (2000), Genetic network inference: from co-expression clustering to reverse engineering, *Bioinformatics*, 16, 707-726.

Dudoit, S., Fridlyand, J. and Speed T., P. (2002), Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97, 77-87.

Dudoit, S., Shaffer, J., P. and Boldrick J., C. (2003), Multiple hypotheses testing in microarray experiments, *Statistical Science*, 18, 71-103.

Dudoit, S. and van der Laan, J., M. (2008), Multiple testing procedures with applications to genomics, *Springer*.

Efron, B. (2003), Robbins, empirical bayes and microarrays, *The Annals of Statistics*, 31, 366-378.

Efron, B. (2004), Large-scale simultaneous hypothesis testing: The choice of null hypothesis, *Journal of the American Statistical Association*, 99, 96-104.

Glazko, G. and Emmert-Streib, F. (2009), Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets, *Bioinformatics*, 25 No. 18, 2348-2354.

Golub, T., R., Slonim, D., K., Tamayo, P., Huard, C., Caasenbeek, 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, 531-537.

Gordon, A., Y. (2007), Unimprovability of the Bonferroni procedure in the class of general step-up multiple testing procedures, *Statistics and Probability Letters*, 7, 117-122.

Gordon, A., Y. (2011), A new optimality property of the Holm step-down procedure, *Statistical Methodology*, 8, 129-135.

Gordon, A., Y. and Salzman, P. (2008), Optimality of the Holm procedure among general step-down multiple testing procedures, *Statistics and Probability Letters*, 78, 1878-1884.

Gordon, A., Y. and Klebanov, L. (2010), On a paradoxical property of the Kolmogorov-Smirnov two-sample test, *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, Vol. 7, 70-74.

Göhlmann, H. and Talloen, W. (2009), Gene eexpression studies using affymetrix microarrays, *Taylor & Francis*.

Hájek, J., Šidák, Z. and Sen, P., K. (1999), Theory of Rank Tests (Second Edition), *Academic Press*.

Holloway, A., Oshlack, A., Diyagama, D., Bowtell, D. and Smyth, G. (2006), Statistical analysis of an RNA titration series evaluates microarray precision and sensitivity on a whole-array basis, *BMC Bioinformatics*, 7, 511.

Holm, S. (1979), A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics*, 6, 65-70.

Kagan, A., M., Linnik, I., V. and Rao, C., R. (1973), Characterization problems in mathematical statistics, *Wiley*.

Klebanov, L., Jordan, C. and Yakovlev, A.(2006), A new type of stochastic dependence revealed in gene expression data, *Statistical Applications in Genetics and Molecular Biology*, 5 issue 1,article 7.

Klebanov, L. and Yakovlev, A. (2007), Diverse correlation structures in gene expression data and their utility in improving statistical inference, *The Annals of Applied Statistics*, 1 No.2, 538-559.

van der Laan, M., J., Dudoit, S. and Pollard K., S. (2004), Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives, *Statistical Applications in Genetics and Molecular Biology*, 3, Article 15.

Lehmann, E., L. and Romano, J., P. (2005), Generalizations of the familywise error rate, *The Annals of Statistics*, 33, 1138-1154.

Lim, J., Kim, J. and Kim, B., S. (2010), An alternative model of type A dependence in a gene set of correlated genes, *Statistical Applications in Genetics and Molecular Biology*, 9, #1.

Logsdon, C., D., Simeone, D., M., Binkley, C., Arumugam T., Greenson, J.,K., Giordano, T.,J., Misek, D.,E., Kuick, R. and Hanash, S. (2003), Molecular profiling of pancreatic adenocarcinoma and chronic pancreatitis identifies multiple genes differentially regulated in pancreatic cancer, *Cancer Research*, 63, 2649-2657.

Massey, F., J. (1950), A Note on the Power of a Non-Parametric Test, *Annals of Mathematical Statistics*, Vol. 21, 440-443.

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

Quackenbush, J. (2001), Computational analysis of microarray data, *Nature Review Genetics*, 2, 418-427.

Robbins, H. (1964), The empirical Bayes approch to statistical decision problems, *The Annals of Mathematical Stastistics*, 35, 1-20.

Roy, S., N. (1957), Union-intersection principle, *Some Aspects of Multivariate Analysis, Wiley, New York*.

Schena, M., Shalon, D., Davis, R., W. and Brown, P., O. (1995), Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270, 467-470.

Sorlie, T. (2001), Gene Expression patterns of breast carcinomas distinguish tumor subclass with clinical implications, *Proceedings of the National Academy of Sciences of USA*, 98, 10869-10874.

Yakovlev, A., Klebanov, L. and Gaile, D. (2013), Statistical Methods for Microarray Data Analysis, *Springer, New York*.

Yang, Y., H., Dudoit, S., Luu, P., Lin, D., M., Peng, V., Ngai, J. and Speed, T., P. (2002), Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research*, 30 (4), e15.

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, Ch., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui Ch.-H., Evans, W.,E., Naeve, C., Wong, L.

and James Downing, J., R. (2002), Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer cell,* 1, 133-143.

Zembutsu, H., Ohnishi, Y., Tsunoda, T., Furukawa, Y., Katagiri, T., Ueyama, Y., Tamaoki, N., Nomura, T., Kitahara, O., Yanagawa, R., Hirata, K. and Nakamura, Y. (2002), Genome-wide cDNA Microarray Screening to Correlate Gene Expression Profiles with Sensitivity of 85 Human Cancer Xenografts to Anticancer Drugs, *Cancer research*, 62, 518527.

Zinger, A., A., Kakosyan, A., V. and Klebanov, L., B. (1989), Characterization of distributions by means of mean values of some statistics in connection with some probability metrics, *Stability Problems for Stochastic Models,* 47-55.