Rohmatul Fajriyah

Microarray data analysis: background correction and differentially expressed genes

DISSERTATION

zur Erlangung des akademischen Grades

Doktors der technischen Wissenschaften

eingereicht an der

Technischen Universität Graz

Betreuer/in:
Univ.-Prof. Mag. Dr.rer.nat István Berkes

Institut für Statistik

Graz, Dezember 2014

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZ-online hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral dissertation.

Datum/Date	Unterschrift/Signature

"The whole of life is just like watching a film. Only it's as though you always get in ten minutes after the big picture has started, and no-one will tell you the plot, so you have to work it out all yourself from the clues."

Terry Pratchett

Abstract

Microarray technologies have been widely used to provide data for research, particularly in life sciences research. These data need to be pre-processed before further analysis is applied, because some steps in producing microarray data contribute to noise. Two of the most advanced platforms in microarray technology are the Affymetrix GeneChips and the Illumina BeadArrays.

We study and compare the performance of the existing convolution models for background correction of the Illumina BeadArrays in the literatures and propose the new approach to adjust the intensity value. Our study shows our model (with the method of moments for the parameter estimation) to be the optimal model for the benchmarking data set with the benchmarking criteria. In the public data sets our proposed models have the best performance, showing only a moderate error in the background correction and in the parametrization.

Further, we generalized the proposed model to the convolution based on the generalized beta distribution. The generalized model facilitates a user in finding a suitable convolution model for the data at hand.

We propose two new tests to compare the mean of two independent samples. The simulation results of the first proposed test show that, it has equal power to the t test. The simulation results of the second proposed test show that, in general, it has a better power than the t test even when the sample size is small. These two proposed tests can be used as the alternative tests of determining the differentially expressed genes under two conditions of investigation in microarray data analysis.

Acknowledgements

I, *sincerely*, would like to thank **Prof. István Berkes**, my PhD supervisor for his help, support, encouragement, and guidance during the preparation of this thesis. I would also like to thank **Prof. Ernst Stadlober** and the staff of the Institute of Statistics of Graz University of Technology for their hospitality and the friendly atmosphere that greatly facilitated my work. Finally, I would like to acknowledge financial support from the **Austrian Science Fund (FWF)**, Project P24302-N18.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

There are various processes in producing data from microarray experiments and each process contributes noise to the data. The noise can be of two types, biological and non-biological. Non-biological noise should be avoided or at least minimized.

Sources for the non-biological noises are, for example, the chip itself, the scanner, or fluctuations in the electrical network. Therefore, it is very important to implement the pre-processing, before further analysis of data is performed. Pre-processing is applied in order to improve the accuracy of the conclusion from the analysis.

Generally, the steps in the pre-processing are image processing, background correction, normalization, and summarization of the probes into a single value [37]. In this thesis, we study the background correction step of the pre-processing. Some researchers ([4], [9], and [49]) believe that the background correction step is the most important step in the pre-processing, because it adjusts ([37, 38]) and provides the true intensity.

To compute the true intensity, researchers have proposed additive, multiplicative and also *additive-muliplicative* error models, see e.g. Huber et al.[36]. In the case of additive models, the underlying distribution is generally chosen as normal (log-normal), exponential, or a gamma-t mixture in the parametric approach ([1], [5], [8], [33], [40–42], and [53, 54]).

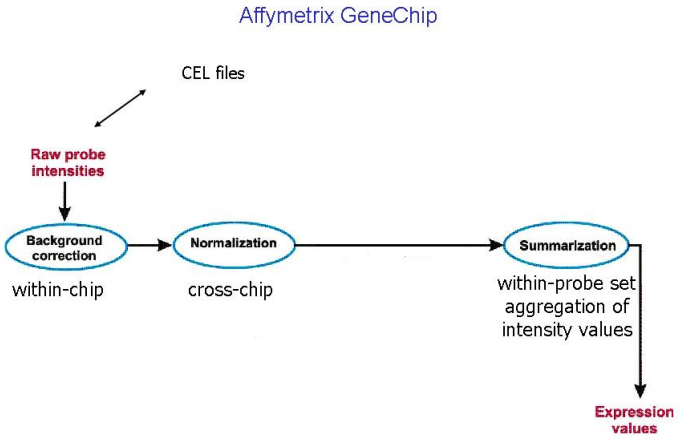


FIGURE 1.1: Pre-processing steps in general, adapted from [38].

Irizarry et al. [40–42] and Bolstad et al. [5], in the Affymetrix platform computed the true intensity based on the convolution model in the background correction step of their robust multi-array average (RMA) pre-processing method. They assumed that the true intensity is exponentially distributed and the background noise is normally distributed.

Plancade et al. [53, 54] showed that the RMA model (in [5] and [40–42]) does not fit the Illumina BeadArrays: using the exponential-normal convolution leads to a large distance between the observed and the modeled intensities. Instead, they proposed the implementation of gamma distribution for the intensity value and normal distribution for the noise.

The simulation study of Plancade et al. [53, 54] showed that the gamma-normal model performs better than the existing exponential-normal convolution model, giving a more accurate and correct fit for the observed intensities in the Illumina BeadArrays.

Using gamma distribution for the intensity values in the Illumina BeadArrays was first suggested by Xie et al. [68].

The studies of Baek et al. [3] (in the background correction of the image processing) and Chen et al. [8] show that the noise distribution is usually skewed in different degrees. Baek et al. [3], in their studies based on the simulated and real data sets, conclude that the gamma distribution is well suited for the noise. It

accounts for the intensities with a positive lower bound and is very flexible in its shape, including asymmetric exponential type and symmetric normal type.

The proposed convolution of exponential-gamma distribution by Chen et al. [8] improves the intensity estimation and the detection of differentially expressed genes in cases where the intensity to noise ratio is large and the noise has a skewed distribution.

In view of the remarks above, it is natural to model both the true intensity and the background noise in the Illumina BeadArrays as gamma distributed. We computed the true intensity value based on the gamma-gamma convolution model of RMA. However, this model did not fit into the Illumina benchmarking data set. Independently, Triche et al. [62] proposed and applied the gamma-gamma model to pre-process the Illumina methylation arrays.

We introduce a new model for background correction in the Illumina BeadArrays where the true intensity value is followed the exponential or gamma distribution and the noise has lognormal distribution. As we will see, this model avoids the difficulties of the gamma-gamma model and has an overall satisfactory performance.

We note that a new method reducing the bias of the maximum likelihood estimator of the shape parameter of the gamma distribution has been proposed by Zhang [70]. Bias is not a problem in our studies, because our samples are very large.

1.2 An overview of gene expression and microarray technology

1.2.1 Gene expression

Biology is the study of living organisms, which are made up of cells. Cells are the fundamental working units of every living organism, where all the instructions needed to direct their activities are contained within the chemical nucleic acid. Nucleic acid is made up of nucleotides, that consist of a nitrogenous base, sugar (pentose) and phosphate.

There are two kinds of nucleic acid, deoxyribose nucleic acid (DNA) and ribose nucleic acid (RNA). There are five different types of base, namely, adenine (A),

cytosine (C), guanine (G), thymine (T), and uracil (U). In nucleic acid, the pentose sugars are deoxyribose and ribose.

DNA is a nucleic acid with a sugar component of deoxyribose and base components of A, C, G and T. The deoxyribose sugar consists of 5 carbon atoms and an oxygen atom in a ring, with the carbon atoms numbered as 5', 4', 3', 2', and 1'. The ' is read as prime, a naming convention that specifies the carbon atoms in the deoxyribose ring, not the carbons of the base.

DNA takes the form of a double helix with two nucleotide chains, containing a linear backbone of sugar (S) and phosphate (P). In this form, the direction of the nucleotides in one strand is the opposite to their direction in the other strand. The ends of DNA strands are called the 5' and 3' ends. This refers to the locations of carbon atoms on the pentose sugars. The structure of DNA can be seen in Figure 1.2.

The double helix is formed due to the hydrogen bonding between base pairs. The bases on one strand are paired with the bases on another strand, according to Watson-Crick base pairing rules, where A specifically pairs with T, and C pairs with G ([2], [13], and [69]). It is repeated millions or billions of times throughout a genome. The particular order of As, Ts, Cs, and Gs dictates whether an organism is human or another species, for example yeast, rice, or fruit fly.

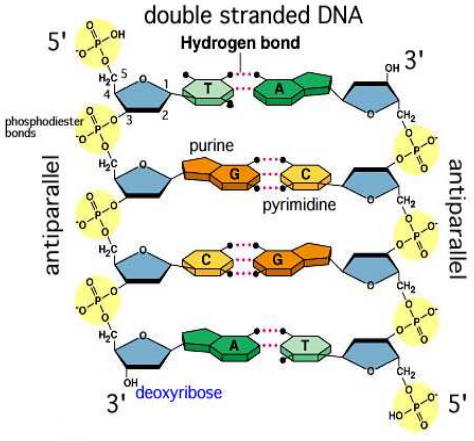


FIGURE 1.2: DNA structure, http://academic.brooklyn.cuny.edu/biology/bio4fv/page/molecular%20biology/dna-structure.html, last retrieved May 6, 2014.

DNA in each human cell is packaged into 46 chromosomes and arranged into 23 pairs. Each chromosome is a physically separate molecule of DNA that ranges in length from about 50 million to 250 million base pairs [2]. Each chromosome

contains many genes, the basic physical and functional units of heredity for an organism [43]. A structure at the end of a chromosome, where there is an area of highly repetitive DNA sequencing is called telomere, see Figure 1.3 and 1.4.

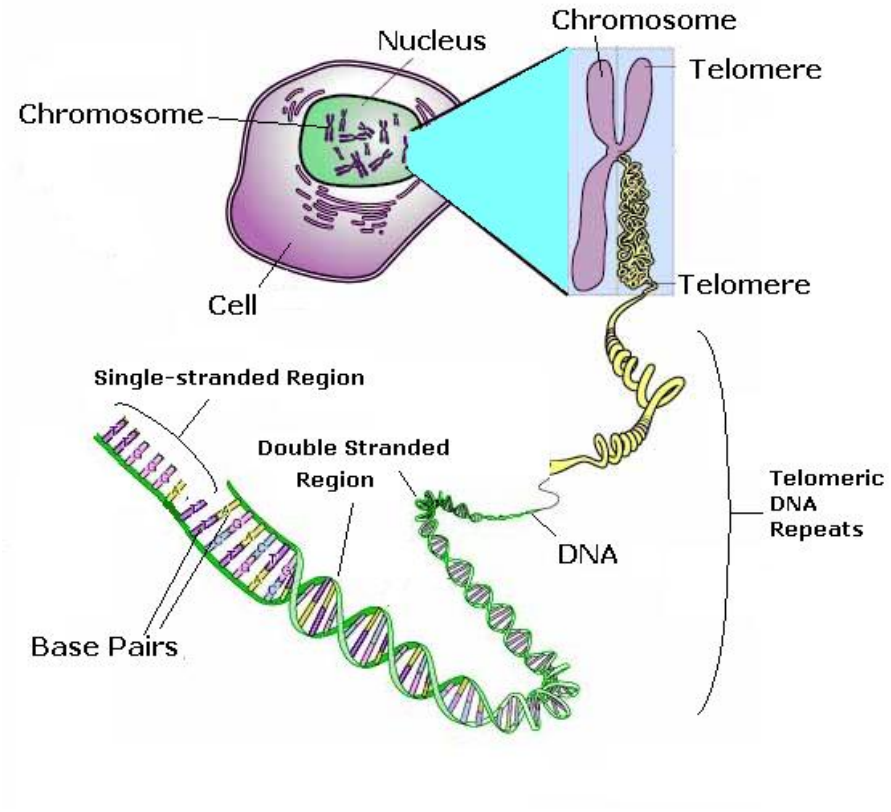


FIGURE 1.3: Cell, Chromosome and DNA, http://www.ch.ic.ac.uk/local/projects/burgoine/origins.txt.html, last retrieved May 6, 2014

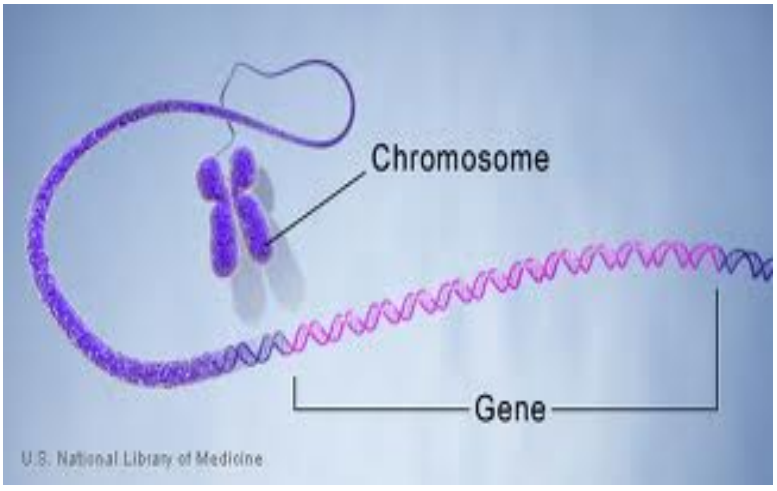


FIGURE 1.4: Gene and DNA, http://www.ghr.nlm.nih.gov/handbook/basics?show=all, last retrieved May 6, 2014.

RNA is a nucleic acid with a sugar component of ribose (has an -OH at the 2' C position, whereas the DNA sugar has an -H at that position) and a base component containing the base uracil instead of thymine. It is a single stranded.

Genes are specific sequences of bases that encode instructions on how to make proteins or an RNA molecule ([2], [13], [43], and [69]). Gene expression is the process whereby a gene transfers its genetic code information from DNA into protein.

Firstly, the DNA double helix splits and develops a condition where one strand of the DNA acts as a template of where the complementary of messenger ribonucleic acid (mRNA) is formed. The mRNA strand then separates. The sequence bases of mRNA are then converted into proteins through the translation step. All the process are formulated in a central dogma of molecular biology [2].

1.2.2 DNA microarray technology

1.2.2.1 DNA microarray technology in general

A DNA microarray is a technology used in molecular biology to monitor gene expression in parallel. Gabig and Wegrzyn [26] define the technology as high density arrays of DNA or oligonucleotide sequences, known as probes, in thousands of features. These probes hybridize the mRNA samples in Watson-Crick base pairing. Because there are probes for each gene, we are able to measure the activity level of genes in a particular sample.

The cells in a human body contain identical genetic material, but the same genes are not active in every cell. To determine which genes are turned on and which are turned off in a given cell, a researcher needs to conduct microarray experiments.

If a particular gene is very active in a given cell, it produces many molecules of mRNA. Therefore, the hybridization process will generate very bright fluorescence. Genes that are less active produce fewer mRNAs and less fluorescence. If there is no fluorescence, none of the messenger molecules have hybridized to the target on the microarray slide, indicating that the gene is inactive. The gene expression is measured by the intensity value of the scanned image of the microarray slide after the steps of hybridization, washing and staining.

Amaratunga and Cabrera [2], Draghici [13], Lee [43], and Zhang [69] explain that the application of microarray technology is related to the post-genomics era, since a GeneChip contains tens of thousands of probes. Because of that, a microarray experiment can monitor the expression pattern of many genes in parallel and therefore researchers can simultaneously investigate many genes and their interaction at once. Previously, this was not possible: the researcher could only monitor a few genes in one experiment.

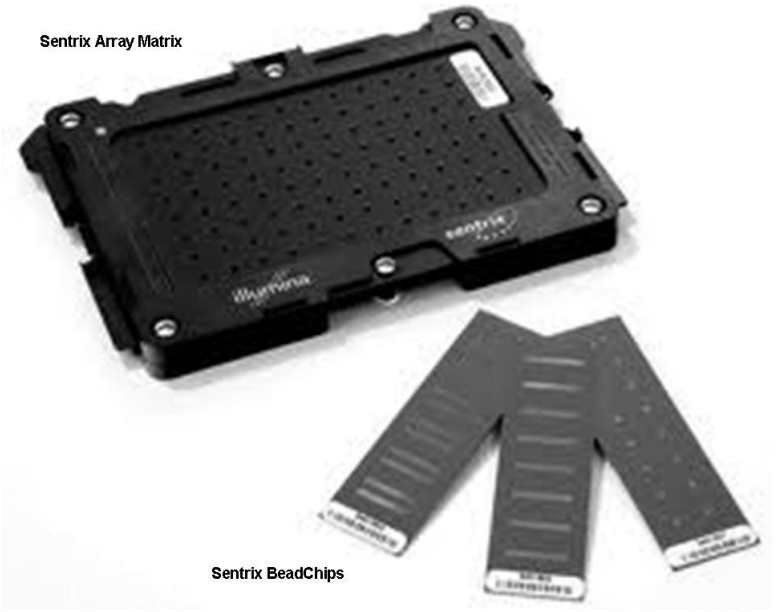


FIGURE 1.5: Illumina platforms, [23]

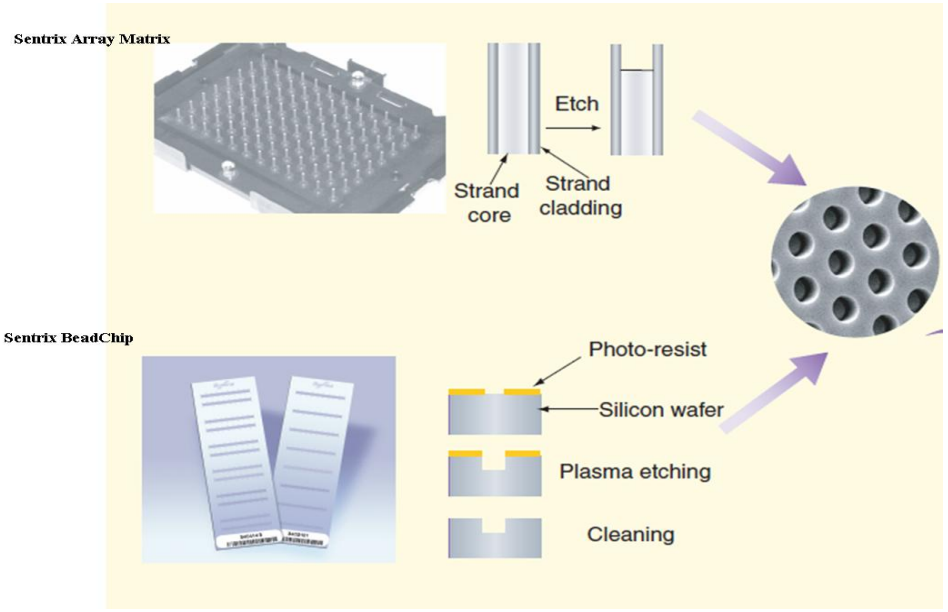


FIGURE 1.6: Pattern substrate of Illumina platform, [61].

1.2.2.2 Illumina BeadArrays technology

Illumina technology is one of the most advanced technologies in analyzing gene expression by microarrays. It can be used to profile partially degraded RNA which is usually found in the FFPE samples, by the cDNA-mediated Annealing, Selection, Extension, and Ligation (DASL) assay method.

The huge amount of available the formalin-fixed, paraffin-embedded (FFPE) data make the Illumina platform very important because of the nature of the DASL assay method, which can deal with the partially degraded RNA to profile the gene expression of the samples.

The Illumina platform has a small feature size, dense features and the ability to analyze multiple samples in parallel. Illumina provides two formats of microarrays ([22], [23], [51] and [61]), the Sentrix® Array Matrix (SAM) and the Sentrix BeadChip (SBC). See Figure 1.5. The pattern substrate can be seen in Figure 1.6.

The Array Matrix arranges fiber optic bundles, each containing 50,000 fibers within a distance of 5-μm, into an Array of Arrays™ format of a 96-well microtiter plate. On one end of the fiber optic bundles, the core of each fiber is etched to form a nanowell for the 3-μm silica beads.

In the BeadChip format, one or more microarrays are arranged on silicon slides that have been processed by micro-electromechanical systems (MEMS) technology to also have nanowells that support the self assembly of beads.

Steemers and Gunderson [61] explain the three parts of the Illumina BeadArrays manufacturing process (see Figure 1.8). The three parts are:

1. The first part is the creation of a master bead pool consisting of 1,536-250,000 different bead types. For quality control, it includes the negative control beads. Oligonucleotide capture probes are immobilized individually by bead type in a bulk process. Each bead type in an array comes from a single immobilization event, reducing array-to-array feature variability. The design of the Illumina bead can be seen in Figure 1.7.

2. The second step is the random self assembly of the master pool of bead types into etched wells on the array substrate, where each bead type is represented on average 30 times - a strategy which provides the statistical accuracy of multiple measurements.

3. The third step is the identification of each bead on the array, through a decoding process. This process provides information of each bead and performs a quality control of the feature in every array.

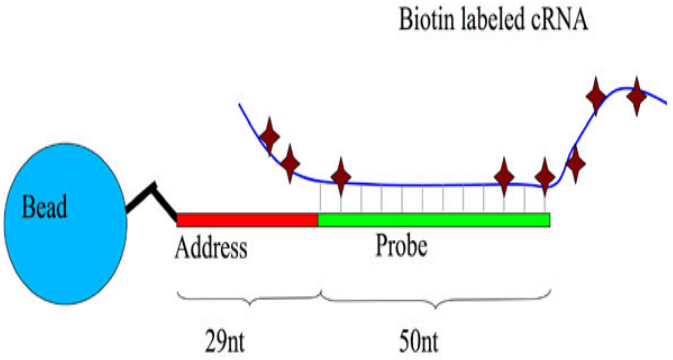


FIGURE 1.7: The design of an Illumina bead. In this figure, the bead is shown to be coated by one oligonucleotide only. In the real bead, it is coated by hundreds of thousands of copies of a specific oligonucleotide. http://bitesizebio.com/articles/how-dna-microarrays-are-built/, last retrieved May 6, 2014.

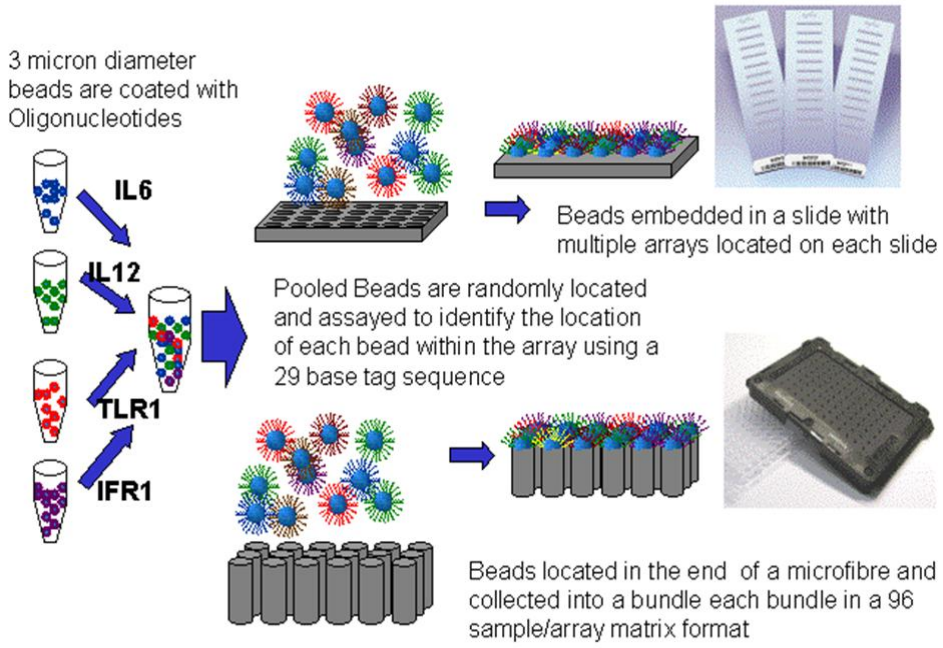


FIGURE 1.8: Production process of Illumina, http://www.ipc.nxgenomics.org/newsletter/no8.htm, last retrieved May 6, 2014.

1.3 Some aspects on microarray data analysis

1.3.1 Benchmarking

The Illumina design provides two probe types: the regular and the control probes. This design is very suitable for studying the probability distribution of both probe types. Therefore, one can apply the convolution model to compute the true intensity value. The availability of benchmarking data in the Illumina platform, the Illumina spike-in, helps researchers to evaluate their proposed method for the Illumina BeadArrays.

We compare performance of models on the Illumina spike-in data set, based on various criteria: root and mean square error (RMSE), L_1 error, Kullback-Leibler (K-L) coefficient, and some adapted criteria from Affycomp [10]. These criteria measure the reproducibility, accuracy, precision, specificity and sensitivity of the expression measure of each model. We then provide a simulation to measure the consistency of errors in background correction and the parametrization. The description and details of these criteria and the simulation can be found in the next sections.

1.3.1.1 Benchmarking data set

The Illumina platform has provided a benchmarking data set, the Illumina spike-in [14]. These spike-in probes are targeting bacterial and viral genes absent from the mouse genome. These were added at specific concentrations on each sample. Therefore the change in expression level of a particular spike between samples is known a priori. The expression levels of the non-spikes should not change between samples.

There are twelve different concentrations of spike: 1000 picomolar (pM), 300 pM, 100 pM, 30 pM, 10 pM, 3 pM, 1 pM, 0.3 pM, 0.1 pM 0.03 pM, 0.01 pM and 0 pM. It was replicated four times. Therefore, there are 48 samples with each sample having regular and control bead-type level probes.

There are approximately 48,000 bead-type level probes for each sample into which 33 spike-in bead-type level probes are added. For the control, there are 1,616

negative control bead-type level probes. These control experiments are the benchmarking data sets of the Illumina platform and are used to compare low-level analysis methods such as in the Affymetrix platform.

1.3.1.2 Benchmarking criteria

Cope et al [10] provides the Affycomp *package* for the benchmarking study. It provides some criteria for the benchmarking study in the Affymetrix platform, but has been used by [68], [8], and [53, 54] in the Illumina platform. We adopt these criteria and the explanation is as follows.

The Affycomp [10] provides fourteen criteria and here we define different ranges for each classification. Cope et al. [10] define the low,medium and high intensities as a nominal concentration less than or equal to 2 pM, a nominal concentration between 4 and 32 pMs, and a nominal concentration greater than or equal to 64 pM respectively.

In the Illumina spike-in data set, the high, medium and low concentrations are defined as a nominal concentration less than or equal to 1 pM, a nominal concentration between 3 and 30 pMs, and a nominal concentration greater than or equal to 100 pM. Instead of using the slope, we used the R^2, as it reflects the best fit for the data: the quantity that measures what percentage of the observed expressions is explained by nominal concentrations.

The criteria 1 to 9 below were computed by the author and criteria 10 to 14 were computed by implementing the *assessSpikeIn2* and the *assessSpikeIn* functions from the Affycomp *package*, with some adjustments. The benchmarking criteria are as follows:

1. *Median SD.* It is believed that the variance of an expression measure across replicate arrays should be low, so the standard deviation (SD) is also low, ideally zero. The median of standard deviations across replicate arrays is chosen to be the measurement due to its robustness.

2. *Null log-fc IQR.* The non spike-in genes should not be differentially expressed across arrays. Therefore, the Inter-quartile range (IQR) of the log-fold-changes of the non spike-in genes is, ideally, zero.

3. *Null log-fc 99.9%*. Same as above but using the 99.9% percentile.

4. *Signal detects R2*. The R-squared (R2) is obtained by regressing expression values on nominal concentrations in the spike-in data. The ideal value of R-squared is 1, because ideally an increment in the nominal concentration is followed by an increment of the expression values, in the same scale.

5. *Low.*R^2. This is obtained from the regression of observed log concentrations on nominal log concentrations for genes with low intensities.

6. *Med.*R^2. This is obtained from the regression of observed log concentrations on nominal log concentrations for genes with medium intensities.

7. *High.*R^2. Same as above but for genes with high intensities.

8. *Obs-intended-fc* R^2. The R^2 that is obtained by regressing observed log-fold-changes against nominal log-fold-changes for the spike-in genes.

9. *Obs-(low)int-fc* R^2. The R^2 that is obtained by regressing observed log-fold-changes against nominal log-fold-changes for the spike in genes with low intensities.

10. *Low AUC*. This is computed as the area under the receiver operator characteristic (ROC) curve (up to 100 false positives) for genes with low intensities, and standardized. Therefore, the optimum value is 1.

11. *Med AUC*. Same as above but for genes with medium intensities.

12. *High AUC*. Same as above but for genes with high intensities.

13. *Weighted avg AUC*. A weighted average of the previous 3 ROC curves with weights related to the amount of data in each classification (low, medium and high).

14. *All AUC*. An AUC for all intensities, 12 arrays.

The criteria above measure

1. accuracy and precision using the squared correlation coefficient and the standard deviation across replicates

2. specificity and sensitivity using the AUC value and the inter-quartile range of log fold-change

1.3.1.3 Affycomp plot

The Affycomp contains some plots that are used as supplemental supports in the process of benchmarking against the spike-in and the dilution data sets. Some of them are as in [10] and http://affycomp.biostat.jhsph.edu.

For the Illumina BeadArrays a slightly different usage is explained as follows:

1. MA plot
   This plot uses 12 arrays representing a single experiment of the Illumina spike-in and the fold changes are generated by comparing the first arrays in the set to each of the others. The spiked-in genes are symbolized by numbers representing the nominal $\log_2$ fold-change for the gene. The non-spike-in genes with observed absolute fold changes larger than 2 are plotted in red. All other probe sets are represented in black.

2. Variance across replicates plot
   Using the benchmarking data set, the variance of an expression measured across replicate array should be low. For each non spiked-in gene in the arrays used in the MA plot, the mean log expression and the observed standard deviation across the replicates are calculated. The resulting scatter plot is smoothed to generate a single curve representing the mean standard deviation as a function of the mean log expression. The standard deviation should be low and independent of the expression level.

3. Observed expression versus nominal expression plot
   In this plot the log observed intensity of spike-in gene is plotted against the log nominal concentration. The average values of observed intensities in each nominal concentration are used to produce a mean curve. Ideally, if the nominal concentration is doubled, the observed intensity as well. Therefore, ideally the observed intensity should be linear in true concentrations with a slope of 1.

4. ROC curve
   Identification of genes which are differentially expressed can be done by filtering the genes using a fold change exceeding a given threshold. An ROC curve offers a graphical representation of both specificity and sensitivity for such a rule. It is constructed by plotting the true positive (TP) rate (sensitivity) against the false positive (FP) rate (1- specificity).

5. Observed fold-change versus nominal fold-change
   The plotting of log fold-change observation and nominal is used for validation of differentially expressed genes.

1.3.1.4 Reproducibility

To assess which models reproduce the best the benchmarking data, two measurements are applied for each j^{th} array (see e.g. Shamilov [57]):

1. Root Mean Square Errors (RMSE), $\text{RMSE}_j = \left(\frac{\sum_{i=1}^{I} (P_{ij} - \widehat{S}_{ij})^2}{I} \right)^{\frac{1}{2}}$
2. Kullback-Leibler (K-L), $\text{K-L}_j = \sum_{i=1}^{I} P_{ij} \log \left(\frac{P_{ij}}{\widehat{S}_{ij}} \right)$,
   where $\widehat{S}_{ij}$ is the background corrected intensity and P_{ij} is the observed intensity.

1.3.2 Differentially expressed genes

Research in the microarray field, although remarkably has solved the problem in life sciences research by providing the huge data of genes which can be investigated at the same time; however, there are still some limitations. For example the replication in the experiments. The rather low sample replications make the decision from the samples not quite conclusive, because the samples are considered not representative enough.

In the beginning, researchers applied one simple assessment, the log-fold-change. This assessment is no longer used, since it does not take into account the variance among samples. Other options implement the nonparametric methods, t test, moderate t test, analysis of variance, or regression test.

In this thesis we propose a new approach to determining the differentially expressed genes, by implementing the cross variance concept.

1.4 Thesis contribution

Although microarray technology has been used to generate accurate, precise and reliable gene expression data [13], there are some outstanding issues relating to data from microarray experiment. Firstly, there are some noise and variation contributions from each step of the microarray fabrication. Secondly, some researchers (e.g. [2], [13], [34], [50], [43], and [69]) believe that once the microarray data are available, the storage, analysis and interpretation of these data present a major challenge due to the massive amount of data generated.

The overall objective of the thesis is to improve the quality of the intensity values of the Illumina BeadArrays and to propose the measurement of differentially expressed genes. The contributions, can be summarized as follows:

1. introducing a new convolution-based model in the background correction step of the Illumina BeadArrays. These results are presented in Chapter 3.
2. providing guidance for users in choosing the best background correction methods for the data at hand. This is presented in Chapter 4.
3. proposing a new approach to determining the differentially expressed genes in microarray experiments by proposing the alternative to the two independent samples t test, which is implementing the cross variance concept. See Chapter 5 for a detailed discussion.

1.5 Structure of the thesis

This thesis has been organized into 6 chapters as follow:

In Chapter 2, the existing background correction methods under the convolution model for the Illumina BeadArrays are explained.

In Chapter 3, the new proposed model and its generalization under the convolution model is introduced.

In Chapter 4, the performance comparison of all models on the benchmarking and the public data sets are described and explained.

In Chapter 5, the proposed alternative to the two independent samples t test (to determine the differentially expressed genes) and its simulation study (to compare its rejection rate and power to the t test) are introduced and explained.

In Chapter 6, the conclusions of the research, the suggestions to the users of pre-processing methods and the future works related to current research are presented.

1.6 Summary of the chapter

In this chapter, the basic concepts of gene expression, microarray technology, and microarray data analysis, the concept of pre-processing steps and differentially expressed genes have been described. The contributions and structure of this thesis have also been outlined. In the next chapter, some background correction convolution-based models will be described.

Chapter 2

The existing background correction under convolution model

2.1 Introduction

The Affymetrix GeneChip is the pioneering and most widely used platform for microarray gene expression experiments. The tools and algorithms used to handle the data are numerous, both free and commercial. Some methods for pre-processing are available. Examples for the pre-processing methods are: MAS5.0 by Affymetrix, multiplicative model based expression index (MMBE) by Li and Wong [45], RMA in Irizarry et al. [40–42] and Bolstad et al. [5], GC-RMA by Wu et al. [67] and maximum likelihood estimation based on the normal-exponential convolution model by Silver et al. [59].

The increasingly popular Illumina BeadArrays is one of the alternative platforms. A few statistical methods have been developed for the BeadArrays data however, there is as of yet, no consensus regarding the pre-processing steps [58].

Ding et al. [12] extended the RMA model by proposing the model-based background correction method (MBCB) and showed that their model leads to a more precise determination of the gene expression and to a better biological interpretation of the Illumina BeadArrays data.

Xie et al. [68] mention that for the background correction step, the Illumina bead studio gives two options (no background correction and background substraction) and the *packages* for the BeadArrays in R provide three options (no background correction, background substraction and RMA background correction).

The studies of Chen et al. [8] and Plancade et al. [53, 54] show that their background correction models are made by adapting the RMA Affymetrix model. As Forcheh et al. [24], pointed out, most preprocessing methods for the Illumina BeadArrays are taken from the Affymetrix microarray platform.

We studied the existing convolution models for background correction of the Illumina BeadArrays in the literature and they are presented in the following sections.

In general, the background correction is applied toward each array, in which there are probes, probesets and genes (terminology for the Affymetrix platform) or bead and bead-type level probes (terminology for the Illumina platform).

In the Illumina platform, each gene is only targeted by one bead-type, which has been represented by about 30 time replications. If we can have a raw benchmarking data set, then it is possible to have all bead-type level probes of the raw data intensities.

The current publicly available benchmarking data set for the Illumina platform is the raw data from the bead studio, which is the average of the bead-type level probes, not background corrected and of unnormalized intensity. Therefore, the background correction in this thesis is applied to the gene intensity in each array.

Suppose we have J arrays and for each array there are I regular genes and M negative control genes. Throughout the thesis, the convolution model is applied for each array j and is represented as follows:

$$P_i = S_i + B_i, \tag{2.1}$$

where P_i, S_i, and B_i are the regular (observed), the true and the noise intensities respectively, $i = 1, ..., I$, $j = 1, ..., J$. For a negative control gene m at array j, $m = 1, 2, ..., M$, the observed intensity denoted by P_{0m} is assumed to be $P_{0m} = B_{0m}$, where B_{0m} is the noise intensity. The P_i and P_{0m} are assumed to be independent.

2.2 Basic concepts

Definition 2.1. Let X and Y be two continuous random variables with density functions $f_1(x)$ and $f_2(y)$ respectively. Assume that both $f_1(x)$ and $f_2(y)$ are defined for all real numbers. Then the *convolution* $f_1 * f_2$ of f_1 and f_2 is the function given by

$$\begin{aligned}(f_1 * f_2)(z) &= \int_{-\infty}^{+\infty} f_1(z-y) f_2(y) dy \\ &= \int_{-\infty}^{+\infty} f_2(z-x) f_1(x) dx. \end{aligned} \tag{2.2}$$

Theorem 2.2. *Let X and Y be two independent random variables with density functions $f_X(x)$ and $f_Y(y)$ respectively defined for all x and y. Then the sum $Z = X + Y$ is a random variable with a density function of $f_Z(z)$, where f_Z is the convolution of f_X and f_Y.*

Definition 2.3. McDonald and Xu [47] define the generalized beta distribution by the probability density function

$$GB_X(x; a, g, c, u, v) = \frac{\mid a \mid x^{au-1}(1-(1-c)(\frac{x}{g})^a)^{v-1}}{g^{au}\mathrm{B}(u,v)(1+c(\frac{x}{g})^a)^{u+v}}; 0 < x^a < \frac{g^a}{1-c}, \tag{2.3}$$

and zero otherwise, with $\mathrm{B}(u, v)$ as the beta function, $0 \leq c \leq 1$, a, g, u and v positive.

Definition 2.4. Let X be a random variable gamma distribution with parameter α and β. The probability density function of X is

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} \exp\{-\beta x\}}{\Gamma(\alpha)}; \quad \alpha, \beta, x > 0. \tag{2.4}$$

Definition 2.5. Let X be a random variable exponential distribution with parameter θ. The probability density function of X is

$$f_X(x; \theta) = \theta \exp\{-\theta x\}; \quad \theta, x > 0. \tag{2.5}$$

Definition 2.6. Let X be a random variable normal distribution with parameter μ and σ^2. The probability density function of X is

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}; \quad \mu_j, x \in \mathbb{R}, \sigma^2 > 0. \tag{2.6}$$

Definition 2.7. Let X be a random variable lognormal distribution with parameter μ and σ^2. The probability density function of X is

$$f_X(x;\mu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\}; \quad \mu_j, x \in \mathbb{R}, \sigma^2 > 0. \tag{2.7}$$

Definition 2.8. A random variable X is confluent hypergeometric distribution $\mathbf{CH}(p,q,s)$ if the probability density function is defined as follows [29]:

$$\mathbf{CH}(p,q,s) = \frac{x^{p-1}(1-x)^{q-1}\exp\{-sx\}}{\mathrm{B}(p,q)_1F_1(p,p+q,-s)}; \quad b,p>0, s\in\mathbb{R}, 0<x<1, \tag{2.8}$$

where $B(p,q)$ is the beta function, $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \frac{(p-1)!(q-1)!}{(p+q-1)!}$,$_1F_1(p,p+q,-s) = \sum_{n=0}^{\infty}\frac{(p)_n}{(p+q)_n n!}(-s)^n$ and $(p)_n = p(p+1)...(p+n-1)$, and $(p)_0 = 1$

2.3 RMA method

The RMA, since its introduction in 2003 [40–42], has gained popularity among bioinformaticians. It has evolved from the exponential-normal convolution to the gamma-normal convolution, from single to two channels and from the Affymetrix to the Illumina platform.

The RMA method was developed for the Affymetrix platform, where the design of arrays differs from the Illumina platform. The Affimetrix paltform has perfect match and mismatch probe design. The RMA uses only the perfect match probe, which is the targeted probe of the intended gene. For those not familiar with the Affymetrix platform, refer to [4], [5], and [40–42] for further information.

In modelling the intensity values, the RMA model ([5], and [40–42]) assumes that the intensity values are affected by the noise of the chip. According to Equation (2.1), in the RMA model $P_i = PM_i$ is the observed probe level intensity of perfect match probes of the i^{th} gene, S_i is the true intensity of the i^{th} gene, with $S_i \sim$

$f_1(s_i;\theta_j) = \text{Exp}(\theta_j), \theta_j, s_i > 0$, and B_i as the background noise of the i^{th} gene with $B_i \sim f_2(b_i;\mu_j,\sigma_j^2) = \mathcal{N}\left(\mu_j,\sigma_j^2\right), \mu_j \in \mathbb{R}, \sigma_j^2, b_i > 0$.

Assuming independence, the joint density of the two-dimensional random variables (S_i, B_i) is

$$f_{S_i,B_i}(s_i,b_i;\mu_j,\sigma_j^2,\theta_j) = \theta_j \exp\{-s_i\theta_j\} f_2\left(b_i;\mu_j,\sigma_j^2\right); s_i, b_i > 0. \tag{2.9}$$

Furthermore, the transformation formula for two-dimensional densities that gives the joint density of S_i and P_i is

$$\begin{aligned}
&f_{S_i,P_i}(s_i,p_i)\\
&= f_{S_i,B_i}(s_i,p_i-s_i;\mu_j,\sigma_j^2,\theta_j)|J|\\
&= f_1(S_i = s_i \mid \theta_j) f_2(P_i - S_i = p_i - s_i \mid \mu_j,\sigma_j^2)\\
&= \theta_j \exp\{-\theta_j s_i\} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left((p_i-s_i)-\mu_j\right)^2 \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left((p_i-s_i)-\mu_j\right)^2 - \theta_j s_i \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left(p_i^2 - 2s_ip_i + s_i^2 - 2(p_i-s_i)\mu_j + \mu_j^2 + 2\theta_j\sigma_j^2 s_i\right) \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left(s_i^2 - 2s_ip_i + 2s_i\mu_j + p_i^2 - 2p_i\mu_j + \mu_j^2 + 2\theta_j\sigma_j^2 s_i\right) \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left(s_i^2 - 2s_i\left(p_i - \mu_j - \theta_j\sigma_j^2\right) + p_i^2 - 2p_i\mu_j + \mu_j^2\right) \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left(\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2 - \right.\right.\\
&\left.\left.\left(p_i - \mu_j - \theta_j\sigma_j^2\right)^2 + p_i^2 - 2p_i\mu_j + \mu_j^2\right) \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left(\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2 - \right.\right.\\
&\left.\left.\left(p_i - \mu_j - \theta_j\sigma_j^2\right)^2 + (p_i-\mu_j)^2\right) \right\}\\
&= \frac{\theta_j}{\sqrt{2\pi}\sigma_j} \exp\left\{ -\frac{1}{2\sigma_j^2}\left(\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2 - \left(-2(p_i-\mu_j)\theta_j\sigma_j^2 + \theta_j^2\sigma_j^4\right)\right) \right\}
\end{aligned}$$

$$= \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} \frac{\exp\left\{-\frac{1}{2\sigma_j^2}\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2\right\}}{\sqrt{2\pi}\sigma_j}$$

$$= \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} f_2\left(s_i; p_i - \mu_j - \theta_j\sigma_j^2, \sigma_j^2\right). \tag{2.10}$$

Therefore,

$$f_{S_i,P_i}(s_i, p_i; \mu_j, \sigma_j^2, \theta_j)$$
$$= \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} f_2\left(s_i; p_i - \mu_j - \theta_j\sigma_j^2, \sigma_j^2\right). \tag{2.11}$$

The marginal density of P_i is

$$f_{P_i}(p_i; \mu_j, \sigma_j^2, \theta_j)$$
$$= \int_0^\infty f_{S_i,P_i}(s_i, p_i; \mu_j, \sigma_j, \theta_j) ds_i$$
$$= \int_0^\infty \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} f_2\left(s_i; p_i - \mu_j - \theta_j\sigma_j^2, \sigma_j^2\right) ds_i$$
$$= \int_0^\infty \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} \frac{\exp\left\{-\frac{1}{2\sigma_j^2}\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2\right\}}{\sqrt{2\pi}\sigma_j} ds_i$$
$$= \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} \int_0^\infty \frac{\exp\left\{-\frac{1}{2\sigma^2}\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2\right\}}{\sqrt{2\pi}\sigma_j} ds_i$$
$$= \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} \left(1 - \int_{-\infty}^0 \frac{\exp\left\{-\frac{1}{2\sigma_j^2}\left(s_i - \left(p_i - \mu_j - \theta_j\sigma_j^2\right)\right)^2\right\}}{\sqrt{2\pi}\sigma_j} ds_i\right)$$
$$= \theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i-\mu_j)\right\} \left(1 - F_2\left(0; p_i - \mu_j - \theta_j\sigma_j^2, \sigma_j^2\right)\right). \tag{2.12}$$

By taking $\mu_{S.P,j} = p_i - \mu_j - \theta_j \sigma_j^2$, then

$$f_P(p_i; \mu_j, \sigma_j^2, \theta_j) = \theta_j \exp\left\{\frac{\theta_j^2 \sigma_j^2}{2} - \theta_j(p_i - \mu_j)\right\}\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right), \quad (2.13)$$

where F_2 is a Gaussian distribution function. Moreover the conditional density of S_i given P_i is

$$\begin{aligned} f_{S_i|P_i}(s_i \mid p_i; \mu_j, \sigma_j^2, \theta_j) &= \frac{f_{S_i,P_i}(s_i, p_i; \mu_j, \sigma_j^2, \theta_j)}{f_{P_i}(p_i; \mu_j, \sigma_j^2, \theta_j)} \\ &= \frac{\theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i - \mu_j)\right\} f_2\left(s_i; \mu_{S.P,j}, \sigma_j^2\right)}{\theta_j \exp\left\{\frac{\theta_j^2\sigma_j^2}{2} - \theta_j(p_i - \mu_j)\right\}\left(1 - F_2(0; \mu_{S.P,j}, \sigma_j^2)\right)} \\ &= \frac{f_2\left(s_i; \mu_{S.P,j}, \sigma_j^2\right)}{\left(1 - F_2(0; \mu_{S.P,j}, \sigma_j^2)\right)}. \end{aligned} \quad (2.14)$$

The background adjusted intensity is computed by the estimated signal given the observed intensity. It is the conditional expectation $E(S_i \mid P_i = p_i)$.

$$\begin{aligned} &E(S_i \mid P_i = p_i) \\ &= \int_0^\infty s_i \frac{f_2\left(s_i; \mu_{S.P,j}, \sigma_j^2\right)}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} ds_i \\ &= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \int_0^\infty s_i f_2\left(s_i; \mu_{S.P,j}, \sigma_j^2\right) ds_i \\ &= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \left(\int_0^\infty s_i \frac{\exp\left\{-\frac{1}{2}\left(\frac{s_i - \mu_{S.P,j}}{\sigma_j}\right)^2\right\}}{\sqrt{2\pi}\sigma_j}\right) ds_i. \end{aligned} \quad (2.15)$$

By taking $\frac{s_i-\mu_{S.P,j}}{\sigma_j} = x_i$ the Equation (2.15) becomes

$$= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \left(\int_{-\frac{\mu_{S.P,j}}{\sigma_j}}^{\infty} \frac{(\sigma_j x_i + \mu_{S.P,j}) \exp\left\{-\frac{x_i^2}{2}\right\}}{\sqrt{2\pi}} dx_i \right)$$

$$= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \left(\frac{\sigma_j}{\sqrt{2\pi}} \int_{-\frac{\mu_{S.P,j}}{\sigma_j}}^{\infty} x_i \exp\left\{-\frac{x_i^2}{2}\right\} dx_i + \right.$$

$$\left. \frac{\mu_{S.P,j}}{\sqrt{2\pi}} \int_{-\frac{\mu_{S.P,j}}{\sigma_j}}^{\infty} \exp\left\{-\frac{x_i^2}{2}\right\} dx_i \right)$$

$$= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \left(\frac{\sigma_j}{\sqrt{2\pi}} \left(\exp\left\{-\frac{x_i^2}{2}\right\} \Big|_{-\frac{\mu_{S.P,j}}{\sigma_j}}^{\infty} \right) + \right.$$

$$\left. \mu_{S.P,j} \int_0^{\infty} \frac{\exp\left\{-\frac{1}{2}\left(\frac{s_{ij}-\mu_{S.P,j}}{\sigma_j}\right)^2\right\}}{\sqrt{2\pi}\sigma_j} ds_i \right)$$

$$= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \left(\frac{\sigma_j}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\mu_{S.P,j}}{\sigma_j}\right)^2\right\} + \right.$$

$$\left. \mu_{S.P,j} \left(1 - \int_{-\infty}^{0} \frac{\exp\left\{-\frac{1}{2}\left(\frac{s_{ij}-\mu_{S.P,j}}{\sigma_j}\right)^2\right\}}{\sqrt{2\pi}\sigma_j} ds_i \right) \right)$$

$$= \frac{1}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)} \left(\sigma_j^2 \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{1}{2}\left(\frac{\mu_{S.P,j}}{\sigma_j}\right)^2\right\} + \right.$$

$$\left. \mu_{S.P,j} \left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right) \right)$$

$$= \frac{\left(\sigma_j^2 f_2\left(0; \mu_{S.P,j}, \sigma_j^2\right) + \mu_{S.P,j}\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)\right)}{\left(1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)\right)}$$
$$= \mu_{S.P,j} + \frac{\sigma_j^2 f_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)}{1 - F_2\left(0; \mu_{S.P,j}, \sigma_j^2\right)}. \tag{2.16}$$

The background adjusted intensity is computed by using Equation (2.16). The parameters are estimated by *ad-hoc* method. We summarize the *ad-hoc* method from Bolstad [4], McGee and Chen [48] and Xie et al. [68] as follows:

1. First, for each array, a non-parametric density function is fitted from the observed intensities, and compute the mode $m_{0,j}$ of this density.

2. Second compute the local mode $m_{1,j}$ from the lower tail of the density (to the left of $m_{0,j}$) and assign $m_{1,j}$ as μ_j. Compute σ_j from the left tail of the density (to the left of $m_{1,j}$) and compute θ_j from the right tail (to the right of $m_{1,j}$).

Note that modelling the noise as a truncated normal variable means that the noise equals 0 with a positive probability p_0, a rather unpleasant feature of the model. As pointed out in [68], however, in practical cases p_0 is rather small, so this problem can be disregarded. To avoid this difficulty, one can model the noise as the absolute value of an $\mathcal{N}(\mu_j, \sigma_j^2)$ variable, which changes the calculations above.

2.4 Exponential-normal MBCB

Xie et al. [68] use the same underlying distributions as the RMA for the background correction. The differences from the RMA ([5] and [40–42]) are as follows :

1. Xie et al. [68] take the $+\infty$ as the upper bound of the integral to compute the marginal density function and the conditional expectation of the true

intensity value. On the other hand, the RMA puts p as the upper bound of the integration.

The background corrected intensity value of Xie et al. [68] is

$$E(S_i \mid P_i = p_i) = \mu_{S.P,j} + \sigma_j \frac{\phi(\frac{\mu_{S.P,j}}{\sigma_j})}{\Phi(\frac{\mu_{S.P,j}}{\sigma_j})}. \tag{2.17}$$

2. Under the convolution model (2.1), where the true intensity value is assumed to be exponentially distributed and the noise is normally distributed, we then need to estimate the parameters θ_j, μ_j, and σ_j^2. Xie et al. [68] offer three parameters estimation methods: method of moments, maximum likelihood, and Bayesian. On the other hand, the RMA applies the *ad-hoc* method.

Ding et al. [12] use the exponential-normal convolution model to correct the background of the Illumina platform by using the Markov chain Monte Carlo simulation.

2.5 Gamma-normal convolution

Plancade et al. [53, 54] introduced gamma-normal convolution to model the background correction of the Illumina BeadArrays. The model is based on the RMA background correction of the Affymetrix GeneChips. Plancade et al. [53, 54] assume that the intensity value is gamma distributed and the noise is normally distributed.

Under the model background correction in (2.1), f_{P_i} is the convolution of f_{S_i} and f_{B_i}. The background corrected intensity $\tilde{S}_i(p_i)$ is computed as the conditional expectation of S_i given $P_i = p_i$:

$$\widetilde{S}_i(p_i) = \frac{\int s_i f^{\text{gam}}_{\alpha_j,\beta_j}(s_i) f^{\text{norm}}_{\mu_j,\sigma_j}(p_i - s_i) ds_i}{\int f^{\text{gam}}_{\alpha_j,\beta_j}(s_i) f^{\text{norm}}_{\mu_j,\sigma_j}(p_i - s_i) ds_i}, \tag{2.18}$$

where

$$f^{\text{gam}}_{\alpha_j,\beta_j}(x_i; \alpha_j, \beta_j) = \frac{\beta_j^{\alpha_j} x_i^{\alpha_j - 1} \exp\{-\beta_j x\}}{\Gamma(\alpha_j)}; \quad \alpha_j, \beta_j, x_i > 0$$

is the gamma density. When S_i is gamma distributed and B_i is normally distributed, the Equation (2.18) does not have analytic expression as it does in

Equations (2.16) and (2.17). Therefore, Plancade et al. [53, 54] implemented the Fast Fourier Transform to estimate the parameters and to correct the background. For the background correction with Fast Fourier Transform, the Equation (2.18) is rewritten as

$$\widetilde{S}_i(p_i \mid \Theta_j) = \frac{\alpha_j\beta_j \int f^{\text{gam}}_{\alpha_j+1,\beta_j}(s_i) f^{\text{norm}}_{\mu_j,\sigma_j}(p_i - s_i) ds_i}{\int f^{\text{gam}}_{\alpha_j,\beta_j}(s_i) f^{\text{norm}}_{\mu_j,\sigma_j}(p_i - s_i) ds_{ij}}, \tag{2.19}$$

where $\Theta_j = (\mu_j, \sigma_j, \alpha_j, \beta_j)$, and $s_i f^{\text{gam}}_{\alpha_j,\beta_j}(s_i) = \alpha_j\beta_j f^{\text{gam}}_{\alpha_j+1,\beta_j}(s_i)$ is valid for every $s_i > 0$.

2.6 Exponential-gamma convolution

Chen et al. [8] proposed in favor of the distribution of the true intensity and its noise, under the convolution model of Equation (2.1), the exponential and gamma distributions respectively. Therefore, $S_i \sim f_1(s_i;\theta_j) = \text{Exp}(\theta_j)$ and $B_i \sim f(b_{ij};\alpha_j,\beta_j) = \text{GAM}(\alpha_j,\beta_j)$, where $s_i, b_i, \theta_j, \alpha_j, \beta_j > 0$.

The corrected background intensity for the proposed model ([8]) is :

$$\widehat{S}_i = p_i - \frac{\int_0^{p_i} b_i^{\alpha_j} \exp\Big\{-(\beta_j - \theta_j)b_i\Big\} db_i}{\int_0^{p_i} b_i^{\alpha_j - 1} \exp\Big\{-(\beta_j - \theta_j)b_i\Big\} db_i}. \tag{2.20}$$

Chen et al. [8] use the method of moment and the maximum likelihood estimation in estimating the parameters in this model. However in the MBCB *package* only the maximum likelihood is applied. In this thesis, the method of moment is also applied in the computation of the background correction.

2.7 Gamma-gamma convolution

The intensity value is modeled similarly to Equation (2.1), where the true intensity, S_i, is gamma distributed, $S_i \sim f_1(s_i;\alpha_{1,j},\beta_{1,j}) = \text{GAM}(\alpha_{1,j},\beta_{1,j})$, $s_i, \alpha_{1,j}, \beta_{1,j} > 0$ and the background, B_i, is gamma distributed, $B_i \sim f_2(b_i;\alpha_2,\beta_{2,j}) = \text{GAM}(\alpha_{2,j},\beta_{2,j})$, where $b_i, \alpha_{2,j}, \beta_{2,j} > 0$.

The joint density function of (S_i, B_i) is

$$\begin{aligned} & f_{S_i,B_i}(s_i, b_i; \alpha_{1,j}, \beta_{1,j}, \alpha_{2,j}, \beta_{2,j}) \\ &= \frac{\beta_{1,j}^{\alpha_{1,j}} s_i^{\alpha_{1,j}-1} \exp\{-\beta_{1,j} s_i\}}{\Gamma(\alpha_{1,j})} \frac{\beta_{2,j}^{\alpha_{2,j}} b_i^{\alpha_{2,j}-1} \exp\{-\beta_{2,j} b_i\}}{\Gamma(\alpha_{2,j})}; s_i, b_i > 0. \end{aligned} \tag{2.21}$$

The corrected background intensity is derived by determining the joint distribution of S_i and P_i. The joint distribution function of S_i and P_i is

$$\begin{aligned} & f_{S_i,P_i}(s_i, p_i) \\ &= f_{S_i,B_i}(s_i, p_i - s_i; \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j})|J| \\ &= f_1(S_i = s_i; \alpha_{1,j}, \beta_{1,j}) f_2(P_i - S_i = p_i - s_i; \alpha_{2,j}, \beta_{2,j}) \\ &= \frac{\beta_{1,j}^{\alpha_{1,j}} s_i^{\alpha_{1,j}-1} \exp\{-\beta_{1,j} s_i\}}{\Gamma(\alpha_{1,j})} \frac{\beta_{2,j}^{\alpha_{2,j}} (p_i - s_i)^{\alpha_{2,j}-1} \exp\{-\beta_{2,j}(p_i - s_i)\}}{\Gamma(\alpha_{2,j})}. \end{aligned} \tag{2.22}$$

The marginal density function of P_i is

$$\begin{aligned} & f_{P_i}(p_i; \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j}) \\ &= \int_0^{p_i} f_{S_i,P_i}(s_i, p_i; \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j}) ds_i \\ &= \int_0^{p_i} \frac{\beta_{1,j}^{\alpha_{1,j}} s_i^{\alpha_{1,j}-1} \exp\{-\beta_{1,j} s_i\}}{\Gamma(\alpha_{1,j})} \frac{\beta_{2,j}^{\alpha_{2,j}} (p_i - s_i)^{\alpha_{2,j}-1} \exp\{-\beta_{2,j}(p_i - s_i)\}}{\Gamma(\alpha_{2,j})} ds_i \\ &= \frac{\beta_{1,j}^{\alpha_{1,j}}}{\Gamma(\alpha_{1,j})} \frac{\beta_{2,j}^{\alpha_{2,j}} \exp\{-\beta_{2,j} p_i\}}{\Gamma(\alpha_2)} \int_0^{p_i} s_i^{(\alpha_{1,j}-1)} (p_i - s_i)^{(\alpha_{2,j}-1)} \exp\{-(\beta_{1,j} - \beta_{2,j}) s_i\} ds_i \\ &= \frac{\beta_{1,j}^{\alpha_{1,j}}}{\Gamma(\alpha_{1,j})} \frac{\beta_{2,j}^{\alpha_{2,j}} \exp\{-\beta_{2,j} p_i\}}{\Gamma(\alpha_{2,j})} p_i^{\alpha_{1,j}+\alpha_{2,j}-1} \times \\ & \int_0^1 u_i^{\alpha_{1,j}-1} (1 - u_i)^{\alpha_{2,j}-1} \exp\{-p_i(\beta_{1,j} - \beta_{2,j}) u_i\} du_i \\ &= \frac{\beta_{1,j}^{\alpha_{1,j}} \beta_{2,j}^{\alpha_{2,j}} \exp\{-\beta_{2,j} p_i\} p_i^{\alpha_{1,j}+\alpha_{2,j}-1}}{\Gamma(\alpha_{1,j} + \alpha_{2,j})} \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j} - \beta_{2,j}))^n}{(\alpha_{1,j} + \alpha_{2,j})_n n!}. \end{aligned} \tag{2.23}$$

The conditional density function of S_i given P_i is

$$
\begin{aligned}
&f_{S_i|P_i}(s_i \mid p_i; \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j}) \\
&= \frac{f_{S_i,P_i}(s_i, p_i; \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j})}{f_P(p_i; \alpha_{1,j}, \alpha_{2,j}, \beta_{1,j}, \beta_{2,j})} \\
&= \frac{\frac{\beta_{1,j}^{\alpha_{1,j}} s_i^{\alpha_{1,j}-1} \exp\{-\beta_{1,j} s_i\}}{\Gamma(\alpha_{1,j})} \frac{\beta_{2,j}^{\alpha_{2,j}} (p_i - s_i)^{\alpha_{2,j}-1} \exp\{-\beta_{2,j}(p_i - s_i)\}}{\Gamma(\alpha_{2,j})}}{\frac{\beta_{1,j}^{\alpha_{1,j}} \beta_{2,j}^{\alpha_{2,j}} \exp\{-\beta_{2,j} p_i\} p_i^{\alpha_{1,j}+\alpha_{2,j}-1}}{\Gamma(\alpha_{1,j}+\alpha_{2,j})} \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}} \\
&= \frac{s_i^{\alpha_{1,j}-1}(p_i - s_i)^{\alpha_{2,j}-1} \exp\{-(\beta_{1,j} - \beta_{2,j}) s_i\}}{\mathrm{B}(\alpha_{1,j}, \alpha_{2,j}) p_i^{\alpha_{1,j}+\alpha_{2,j}-1} \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}}.
\end{aligned} \tag{2.24}
$$

The corrected background intensity given the observed intensities is the conditional expectation $E(S_i \mid P_i = p_i)$. This is computed as follows:

$$
\begin{aligned}
E(S_i \mid P_i = p_i) &= \int_0^{p_i} s_i \frac{s_i^{\alpha_{1,j}-1}(p_i - s_i)^{\alpha_{2,j}-1} \exp\{-(\beta_{1,j} - \beta_{2,j}) s_i\}}{\mathrm{B}(\alpha_{1,j}, \alpha_{2,j}) p_i^{\alpha_{1,j}+\alpha_{2,j}-1} \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}} ds_i \\
&= \frac{1}{\mathrm{B}(\alpha_{1,j}, \alpha_{2,j}) p_i^{\alpha_{1,j}+\alpha_{2,j}-1} \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}} \times \\
&\quad \int_0^{p_i} s_i^{\alpha_{1,j}} (p_i - s_i)^{\alpha_{2,j}-1} \exp\{-(\beta_{1,j} - \beta_{2,j}) s_i\} ds_i \\
&= \frac{p_i}{\mathrm{B}(\alpha_{1,j}, \alpha_{2,j}) \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}} \times \\
&\quad \int_0^1 u_i^{\alpha_{1,j}} (1 - u_i)^{\alpha_{2,j}-1} \exp\{-p_i(\beta_{1,j} - \beta_{2,j}) u_i\} du_i \\
&= \frac{p_i}{\mathrm{B}(\alpha_{1,j}, \alpha_{2,j}) \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}} \times \\
&\quad \mathrm{B}(\alpha_{1,j} + 1, \alpha_{2,j}) \sum_{n=0}^{\infty} \frac{(\alpha_{1,j} + 1)_n (-p_i(\beta_{1,j} - \beta_{2,j}))^n}{(\alpha_{1,j} + \alpha_{2,j} + 1)_n n!} \\
&= \frac{p_i \mathrm{B}(\alpha_{1,j} + 1, \alpha_{2,j}) \sum_{n=0}^{\infty} \frac{(\alpha_{1,j}+1)_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j}+1)_n n!}}{\mathrm{B}(\alpha_{1,j}, \alpha_{2,j}) \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}}
\end{aligned}
$$

$$= \frac{p_i \alpha_{1,j} \sum_{n=0}^{\infty} \frac{(\alpha_{1,j}+1)_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j}+1)_n n!}}{(\alpha_{1,j}+\alpha_{2,j}) \sum_{n=0}^{\infty} \frac{(\alpha_{1,j})_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(\alpha_{1,j}+\alpha_{2,j})_n n!}}. \tag{2.25}$$

In the case that α_j values are differents, where the values of $\beta_{1,j} = \beta_{2,j} = \beta_j$, then the background adjusted intensity is computed by using the Equation (2.26).

$$= \frac{p_i \alpha_{1,j}}{\alpha_{1,j} + \alpha_{2,j}}. \tag{2.26}$$

If β_j values are differents, where the values of $\alpha_{1,j} = \alpha_{2,j} = \alpha_j$, then the background adjusted intensity is as follows

$$= \frac{p_i \sum_{n=0}^{\infty} \frac{(\alpha_j+1)_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(2\alpha_j+1)_n n!}}{2 \sum_{n=0}^{\infty} \frac{(\alpha_j)_n (-p_i(\beta_{1,j}-\beta_{2,j}))^n}{(2\alpha_j)_n n!}}. \tag{2.27}$$

In extreme cases, where the values of $\alpha_{1,j} = \alpha_{2,j}$ and $\beta_{1,j} = \beta_{2,j}$, then the background adjusted intensity is

$$= \frac{p_i}{2}. \tag{2.28}$$

This gamma-gamma convolution model does not fit to the benchmarking data set therefore, we do not use this model in the comparison. We describe the model here, as it was independently proposed and successfully used by Triche et al. [62] for the Illumina methylation arrays.

2.8 Summary of the chapter

In this chapter, the existing background correction convolution-based methods to adjust the intensity value are described in detail and they are including a gamma-gamma convolution model. The next chapter will introduce the proposed convolution model and its generalization.

Chapter 3

The Proposed models

3.1 Introduction

In this chapter, we present the results of the two proposed models of convolution (the exponential-lognormal and the gamma-lognormal) and their generalization by considering two possible types of noise distribution: symmetrical or skewed.

The description of the exponential-lognormal convolution is in Section 3.2 and the gamma-lognormal convolution is in Section 3.3. In Section 3.4 we describe the generalized models of convolution for the background correction of the Illumina BeadArrays: the generalized beta (Section 3.4.2) and the generalized beta-normal distribution convolution (Section 3.4.3).

In each section, the formula to estimate the true intensity value is derived and the methods to estimate the parameters are explained.

3.2 Exponential-lognormal convolution, [19]

3.2.1 Background correction

Consider the model (2.1), when the true intensity S_i is exponentially distributed, i.e. $S_i \sim f_1(s_i;\theta_j) = \theta_j \exp\{-\theta_j s_i\}; \theta_j, s_i > 0$, and the background noise B_i is lognormally distributed, $B_i \sim f_2(b_i;\mu_j,\sigma_j^2) = \frac{\exp\left\{-\frac{(\ln b_i-\mu_j)^2}{2\sigma_j^2}\right\}}{b_i\sigma_j\sqrt{2\pi}}; \mu_j \in \mathbb{R}, \sigma_j^2, b_i > 0.$

The joint density function of S_i and B_i equals

$$f_{S_i,B_i}(s_i;b_i)=\theta_j\exp\Big\{-\theta_j s_i\Big\}\frac{\exp\Big\{-\frac{(\ln b_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{b_i\sigma_j\sqrt{2\pi}}, \tag{3.1}$$

and thus the joint density function of S_i and P_i is

$$f_{S_i,P_i}(s_i;p_i)=\theta_j\exp\Big\{-\theta_j s_i\Big\}\frac{\exp\Big\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\Big\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}};\quad s_i<p_i. \tag{3.2}$$

Consequently, the marginal density function of P_i equals

$$\begin{aligned}f_{P_i}(p_i;\theta_j,\mu_j,\sigma_j^2)&=\int_0^{p_i} f_{S_i,P_i}(s_i;p_i)ds_i\\&=\int_0^{p_i}\theta_j e^{-\theta_j s_i}\frac{\exp\Big\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\Big\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}}ds_i.\end{aligned} \tag{3.3}$$

Using the substitution $\ln(p_i-s_i)=z_i$ then the equation (3.3) can be written as follows:

$$\begin{aligned}f_{P_i}(p_i)&=\int_{-\infty}^{\ln p_i}\frac{\theta_j\exp\Big\{-\theta_j(p_i-\exp\{z_i\})\Big\}\exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{\sigma_j\sqrt{2\pi}}dz_i\\&=\frac{\theta_j\exp\Big\{-\theta_j p_i\Big\}}{\sigma_j\sqrt{2\pi}}\int_{-\infty}^{\ln p_i}\exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}\sum_{k=0}^{\infty}\frac{\theta_j^k\exp\Big\{kz_i\Big\}}{k!}dz_i\\&=\frac{\theta_j\exp\Big\{-\theta_j p_i\Big\}}{\sigma_j\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{\theta_j^k}{k!}\int_{-\infty}^{\ln p_i}\exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}+kz_i\Big\}dz_i\\&=\frac{\theta_j e^{-\theta_j p_i}}{\sigma_j\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{\theta_j^k\exp\Big\{k(\mu_j+\frac{k\sigma_j^2}{2})\Big\}}{k!}\int_{-\infty}^{\ln p_i}\exp\Big\{-\frac{(z_i-(\mu_j+k\sigma_j^2))^2}{2\sigma_j^2}\Big\}dz_i\\&=\theta_j\exp\Big\{-\theta_j p_i\Big\}\sum_{k=0}^{\infty}\frac{\theta_j^k}{k!}\exp\Big\{k(\mu_j+\frac{k}{2}\sigma_j^2)\Big\}\Phi\left(\frac{\ln p_i-(\mu_j+k\sigma_j^2)}{\sigma_j}\right)\\&=\theta_j\exp\Big\{-\theta_j p_i\Big\}C_{1,j},\end{aligned} \tag{3.4}$$

where

$$C_{1,j}=\sum_{k=0}^{\infty}\frac{\theta_j^k}{k!}\exp\Big\{k(\mu_j+\frac{k}{2}\sigma_j^2)\}\Phi\Big(\frac{\ln p_i-(\mu_j+k\sigma_j^2)}{\sigma_j}\Big).$$

The conditional density function of S_i given $P_i = p_i$ is now obtained as

$$\begin{aligned} f_{S_i|P_i}(s_i \mid p_i) &= \frac{f_{S_i,P_i}(s_i, p_i)}{f_{P_i}(p_i)} \\ &= \frac{\theta_j \exp\Big\{-\theta_j s_i\Big\} \frac{\exp\Big\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\Big\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}}}{\theta_j \exp\Big\{-\theta_j p_i\Big\} C_{1,j}} \\ &= \frac{\exp\Big\{\theta_j(p_i - s_i)\Big\} \exp\Big\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\Big\}}{C_{1,j}(p_i - s_i)\sigma_j\sqrt{2\pi}}. \end{aligned} \tag{3.5}$$

The true intensity value is computed by the expectation of the conditional density function in (3.5). It is computed as follows:

$$\begin{aligned} E(S_i \mid P_i = p_i) &= \int_0^{p_i} s_i f(s_i \mid p_i) ds_i \\ &= \frac{\exp\Big\{\theta_j p_i\Big\}}{C_{1,j}} \int_0^{p_i} \frac{s_i \exp\Big\{-\theta_j s_i\Big\} \exp\Big\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\Big\}}{(p_i - s_i)\sigma_j\sqrt{2\pi}} ds_i. \end{aligned} \tag{3.6}$$

Using the substitution $\ln(p_i - s_i) = z_i$, we see that the conditional mean in (3.6) equals

$$\begin{aligned} &= \frac{p_i}{C_{1,j}} \int_{-\infty}^{\ln p_i} \frac{(1 - \frac{\exp\{z_i\}}{p_i}) \exp\Big\{\theta_j \exp\{z_i\}\Big\} \exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{\sigma_j\sqrt{2\pi}} dz_i \\ &= \frac{p_i}{C_{1,j}} \Bigg[\int_{-\infty}^{\ln p_i} \frac{\exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{\sigma_j\sqrt{2\pi}} \exp\Big\{\theta_j \exp\{z_i\}\Big\} dz_i - \\ &\int_{-\infty}^{\ln p_i} \frac{\exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{\sigma_j\sqrt{2\pi}} \frac{\exp\{z_i\}}{p_i} \exp\Big\{\theta_j \exp\{z_i\}\Big\} dz_i \Bigg] \\ &= \frac{p_i}{C_{1,j}} \Bigg[C_{1,j} - \int_{-\infty}^{\ln p_i} \frac{\exp\Big\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{\sigma_j\sqrt{2\pi}} \frac{\exp\{z_i\}}{p_i} \exp\Big\{\theta_j \exp\{z_i\}\Big\} dz_i \Bigg] \\ &= p_i - \frac{e^{\mu_j + \frac{\sigma_j^2}{2}}}{C_{1,j}} \int_{-\infty}^{\ln p_i} \frac{e^{-\frac{(z_i-(\mu_j+\sigma_j^2))^2}{2\sigma_j^2}} e^{\theta_j e^{z_i}}}{\sigma_j\sqrt{2\pi}} dz_i \end{aligned}$$

$$= p_i - \frac{\exp\left\{\mu_j + \frac{\sigma_j^2}{2}\right\}}{C_{1,j}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left\{k(\mu_j + \frac{k+2}{2}\sigma_j^2)\right\} \times$$
$$\int_{-\infty}^{\ln p_i} \frac{\exp\left\{ - \frac{(z_i - (\mu_j + (k+1)\sigma_j^2))^2}{2\sigma_j^2}\right\}}{\sigma_j\sqrt{2\pi}} dz_i$$
$$= p_i - \frac{\exp\left\{\mu_j + \frac{\sigma_j^2}{2}\right\}}{C_{1,j}} \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left\{k(\mu_j + \frac{k+2}{2}\sigma_j^2)\right\} \Phi\Big(\frac{\ln p_i - (\mu_j + (k+1)\sigma_j^2)}{\sigma_j}\Big)$$
$$= p_i - \frac{\exp\left\{\mu_j + \frac{\sigma_j^2}{2}\right\} C_{2,j}}{C_{1,j}}, \tag{3.7}$$

where

$$C_{2,j} = \sum_{k=0}^{\infty} \frac{\theta_j^k}{k!} \exp\left\{k(\mu_j + \frac{k+2}{2}\sigma_j^2)\right\} \Phi\Big(\frac{\ln p_{ij} - (\mu_j + (k+1)\sigma_j^2)}{\sigma_j}\Big).$$

3.2.2 Parameter estimation

To estimate the parameters θ_j, μ_j, and σ_j, $j = 1, 2, .., J$ in our exponential-lognormal model we can use various methods.

1. Maximum likelihood estimation (MLE)
   This is implemented by applying the *optim* function in R to maximize the log-likelihood function of j^{th} array

$$= \sum_{i=1}^{I} \Big\{ \ln(\theta_j) - \theta_j p_i + \ln(C_{1,j}) \Big\} + \sum_{m=1}^{M} \Big\{ - \frac{(\log b_{0m} - \mu_j)^2}{2\sigma_j^2} - \log(b_{0m})$$
$$- \log(\sigma_j) - \frac{\log(2\pi)}{2} \Big\} \tag{3.8}$$

   where p_i and b_{0m} are the observed values of P_i and B_{0m}. Note that $C_{1,j}$ in the log likelihood function is defined by the infinite series at the end of the previous section. However, the terms of this infinite series decrease very rapidly and thus we can cut off the series at a proper index K, making it suitable for R computation. K is chosen by using the criteria $\big|C_{1,j,K+1} - C_{1,j,K} < 0.001\big|$

2. Method of moments
   The implementation of this method at the j^{th} array is applied by recalling that the first two moment estimators of the exponential distribution are

$\frac{1}{I}\sum_{i=1}^{I} S_i = \text{mean}(S_j) = \theta_j$ and $\frac{1}{I}\sum_{i=1}^{I} S_i^2 = \theta_j + \theta_j^2$. On the other hand, the first two moment estimators of the lognormal distribution are $\frac{1}{M}\sum_{m=1}^{M} \log(B_{0mj}) = \text{mean}(\log(\mathbf{B}_{0j})) = \mu_j$ and $\frac{1}{M}\sum_{m=1}^{M} \log(B_{0mj})^2 = \sigma_j^2 + \mu_j^2$. Therefore, when considering Equation (2.1)

(a) θ_j is estimated by $\text{mean}(\mathbf{S}_j) = \text{mean}(\mathbf{P}_j) - \text{mean}(\mathbf{B}_{0j})$

(b) μ_j and σ_j are estimated by $\text{mean}(\log \mathbf{B}_{0j})$ and $\sqrt{\text{var}(\log(\mathbf{B}_{0j}))}$

3. Plug-in
   The plug-in estimate is implemented by estimating

   (a) θ_j from the regular bead-type level probes intensities at the j^{th} array $\mathbf{P}_j$ through MLE, and

   (b) μ_j and σ_j from the negative control bead-type level probes intensities at the j^{th} array $\mathbf{B}_{0j}$ through MLE

3.3 Gamma-lognormal convolution, [19]

3.3.1 Background correction

Consider now model (2.1), when the true intensity S_i is assumed to be gamma distributed, $S_i \sim f_1(s_i;\alpha_j,\beta_j) = \frac{\beta_j^{\alpha_j} s_i^{\alpha_j-1} \exp\left\{-s_i\beta_j\right\}}{\Gamma(\alpha_j)}; \alpha_j, \beta_j, s_i > 0$, and the background noise B_i is lognormally distributed, i.e. $B_i \sim f_2(b_i;\mu,\sigma^2) = \frac{\exp\left\{-\frac{(\ln b_i-\mu_j)^2}{2\sigma_j^2}\right\}}{b_i\sigma_j\sqrt{2\pi}}$; $\mu_j \in \mathbb{R}, \sigma_j^2 > 0$. The joint density function of S_i and B_i is

$$f_{S_i,B_i}(s_i,b_i) = \frac{\beta_j^{\alpha_j} s_i^{\alpha_j-1} \exp\left\{-s_i\beta_j\right\}}{\Gamma(\alpha_j)} \frac{\exp\left\{-\frac{(\ln b_i-\mu_j)^2}{2\sigma_j^2}\right\}}{b_i\sigma_j\sqrt{2\pi}}, \tag{3.9}$$

and therefore, the joint density function of S_i and P_i is

$$f_{S_i,P_i}(s_i,p_i) = \frac{\beta_j^{\alpha_j} s_i^{\alpha_j-1} \exp\left\{-s_i\beta_j\right\}}{\Gamma(\alpha_j)} \frac{\exp\left\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\right\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}}. \tag{3.10}$$

Hence the marginal density function of P_i is obtained as

$$f_{P_i}(p_i) = \int_0^{p_i} f_{S_i,P_i}(s_i, p_i)ds_i$$
$$= \int_0^{p_i} \frac{\beta_j^{\alpha_j} s_i^{\alpha_j-1} \exp\Big\{ - s_i\beta_j\Big\}}{\Gamma(\alpha_j)} \frac{\exp\Big\{ - \frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\Big\}}{(p_i - s_i)\sigma_j\sqrt{2\pi}} ds_i. \tag{3.11}$$

Using the substitution $\ln(p_i - s_i) = z_i$, we get

$$\begin{aligned}
&f_{P_i}(p_i)\\
&= \int_\infty^{\ln p_i} \frac{\beta_j^{\alpha_j} p_i^{\alpha_j-1}(1 - \frac{\exp\{z_i\}}{p_i})^{\alpha_j-1} \exp\Big\{ - p_i\beta_j\Big\} \exp\Big\{ \exp\{z_i\beta_j\}\Big\} \exp\Big\{ - \frac{(z_i-\mu_j)^2}{2\sigma_j^2}\Big\}}{\Gamma(\alpha_j)\sigma_j\sqrt{2\pi}} dz_i\\
&= \frac{\beta_j^{\alpha_j} p_i^{\alpha_j-1} \exp\Big\{ - p_i\beta_j\Big\}}{\Gamma(\alpha_j)\sigma_j\sqrt{2\pi}} \times\\
&\int_\infty^{\ln p_i} \exp\Big\{ - \frac{(z_i - \mu_j)^2}{2\sigma_j^2}\Big\}(1 - \frac{\exp\{z_i\}}{p_i})^{\alpha_j-1} \exp\Big\{ \exp\{z_i\beta_j\}\Big\} dz_i\\
&= \frac{\beta_j^{\alpha_j} p_i^{\alpha_j-1} \exp\Big\{ - p_i\beta_j\Big\}}{\Gamma(\alpha_j)} \Bigg[\sum_{k=0}^{\infty} \frac{(-1)^k \binom{\alpha_j-1}{k}}{p_i^k} \times\\
&\Bigg[\sum_{n=0}^{\infty} \frac{\beta_j^n \exp\Big\{(k+n)(\mu_j + (k+n)\frac{\sigma_j^2}{2})\Big\}\Phi\Big(\frac{\ln p_i - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\Big)}{n!} \Bigg]\Bigg]\\
&= \frac{\beta_j^{\alpha_j} p_i^{\alpha_j-1} \exp\Big\{ - p_i\beta_j\Big\} C_{3,j}}{\Gamma(\alpha_j)},
\end{aligned} \tag{3.12}$$

where

$$C_{3,j} = \sum_{k=0}^{\infty}\sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j-1}{k}}{p_i^k} \frac{\beta_j^n \exp\Big\{(k+n)(\mu_j + (k+n)\frac{\sigma_j^2}{2})\Big\}\Phi\Big(\frac{\ln p_i - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\Big)}{n!}.$$

The conditional density function of S_i under $P_i = p_i$ is now obtained as

$$\begin{aligned} f_{S_i|P_i}(s_i \mid p_i) &= \frac{f_{S_i,P_i}(s_i,p_i)}{f_{P_i}(p_i)} \\ &= \frac{\frac{\beta_j^{\alpha_j} s_i^{\alpha_j-1} \exp\left\{-\beta_j s_i\right\}}{\Gamma(\alpha_j)} \frac{\exp\left\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\right\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}}}{\frac{\beta_j^{\alpha_j} p_i^{\alpha_j-1} \exp\left\{-p_i\beta_j\right\} C_{3,j}}{\Gamma(\alpha_j)}} \\ &= \frac{\exp\left\{p_i\beta_j\right\}}{C_{3,j} p_i^{\alpha_j-1}} \frac{s_i^{\alpha_j-1} \exp\left\{-s_i\beta_j\right\} \exp\left\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\right\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}}. \end{aligned} \tag{3.13}$$

The true intensity value is computed by the expectation of the conditional density function in (3.13) It is computed as follows:

$$\begin{aligned} &E(S_i \mid P_i = p_i) \\ &= \int_0^{p_i} s_i f(s_i \mid p_i) ds_i \\ &= \frac{\exp\left\{p_i\beta_j\right\}}{C_{3,j} p_i^{\alpha_j-1}} \int_0^{p_i} \frac{s_i^{\alpha_j} \exp\left\{-s_i\beta_j\right\} \exp\left\{-\frac{(\ln(p_i-s_i)-\mu_j)^2}{2\sigma_j^2}\right\}}{(p_i-s_i)\sigma_j\sqrt{2\pi}} ds_i. \end{aligned} \tag{3.14}$$

Substituting $\ln(p_i - s_i) = z_i$ the equation (3.14) becomes

$$\begin{aligned} &= \frac{p_i}{C_{3,j}} \int_{-\infty}^{\ln p_i} \frac{\exp\left\{-\frac{(z_i-\mu_j)^2}{2\sigma_j^2}\right\}}{\sigma_j\sqrt{2\pi}} (1 - \frac{\exp\{z_i\}}{p_i})^{\alpha} \exp\left\{\exp\left\{z_i\beta_{\,i}\right\}\right\} dz_i \\ &= \frac{p_i}{C_{3,j}} \sum_{k=0}^{\infty} \frac{(-1)^k \binom{\alpha_j}{k}}{p_i^k} \left[\sum_{n=0}^{\infty} \frac{\beta_j^n \exp\left\{(k+n)(\mu_j + (k+n)\frac{\sigma_j^2}{2})\right\} \Phi\left(\frac{\ln p_i - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\right)}{n!}\right] \\ &= \frac{p_i C_{4,j}}{C_{3,j}}, \end{aligned} \tag{3.15}$$

where

$$C_{4,j} = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{(-1)^k \binom{\alpha_j}{k}}{p_i^k} \frac{\beta_j^n \exp\left\{(k+n)(\mu_j + (k+n)\frac{\sigma_j^2}{2})\right\} \Phi\left(\frac{\ln p_i - (\mu_j + (k+n)\sigma_j^2)}{\sigma_j}\right)}{n!}.$$

3.3.2 Parameter estimation

To estimate the parameters α_j, β_j, μ_j, and σ_j in (3.14), j^{th} array, we can use one of the following methods.

1. Maximum likelihood (MLE)
   This is implemented by applying the *optim* function in R to maximize the log-likelihood function of j^{th} array

$$= \sum_{i=1}^{I} \Big\{ \log(C_{3,j} + (\alpha_j - 1)\log(p_i) - p_i\beta_j - \alpha_j \log(\beta_j) - \log(\Gamma(\alpha_j)) \Big\}$$
$$+ \sum_{m=1}^{M} \Big\{ - \frac{(\log b_{0m} - \mu_j)^2}{2\sigma_j^2} - \log(b_{0m}) - \log(\sigma_j) - \frac{\log(2\pi)}{2} \Big\} \qquad (3.16)$$

   Similar to exponential-lognormal model, in the computation of $C_{3,j}$ K is chosen by using the criteria $\mid C_{3,j,K+1} - C_{3,j,K} \mid < 0.002$

2. Method of moments
   The implementation of this method at the j^{th} array is applied by recalling that the first two moment estimators of gamma distribution are $\frac{1}{I}\sum_{i=1}^{I} S_i = \text{mean}(S_j) = \frac{\alpha_j}{\beta_j}$ and $\frac{1}{I}\sum_{i=1}^{I} S_i^2 = \frac{\alpha_j}{\beta_j^2} + \frac{\alpha_j^2}{\beta_j^2}$. On the other hand, the first two moment estimators of lognormal distribution are $\frac{1}{K}\sum_{m=1}^{M} \log(B_{0mj}) = \text{mean}(\log(\mathbf{B}_{0j})) = \mu_j$ and $\frac{1}{K}\sum_{m=1}^{M} \log(B_{0m})^2 = \sigma_j^2 + \mu_j^2$. Therefore, by considering Equation (2.1)

   (a) β_j and α_j are estimated by $\hat{\beta}_j = \frac{\text{mean}(\mathbf{S}_j)}{\text{var}(\mathbf{S}_j)} = \frac{\text{mean}(\mathbf{P}_j) - \text{mean}(\mathbf{B}_{0j})}{\text{var}(\mathbf{P}_j) - \text{var}(\mathbf{B}_{0j})}$ and $\text{mean}(\mathbf{S}_j)\hat{\beta}_j = (\text{mean}(\mathbf{P}_j) - \text{mean}(\mathbf{B}_{0j}))\hat{\beta}_j$.

   (b) μ_j and σ_j are estimated by $\text{mean}(\log \mathbf{B}_{0j})$ and $\sqrt{\text{var}(\log(\mathbf{B}_{0j}))}$

3. Plug-in
   From Equation (2.1), it is known that $\mathbf{P}_j$ and $\mathbf{B}_{0j}$ are the observed intensities. Therefore, the plug-in estimation is implemented by

   (a) computing α_j and β_j from the regular bead-type level probes intensities at the j^{th} array $\mathbf{P}_j$ through MLE, and

   (b) computing μ_j and σ_j from the negative control bead-type level probes intensities at the j^{th} array $\mathbf{B}_{0j}$ through MLE

3.4 Generalized convolution models, [20]

3.4.1 Motivation

Microarray data come from many steps of production and have been known to contain noise. Pre-processing is implemented to reduce the noise, wherein the background is corrected. Many Illumina BeadArrays users had applied the convolution model, a model which had been adapted from when it was first developed on the Affymetrix platform, to adjust the intensity which provides the corrected background intensity value.

There are a few models currently available to adjust the intensity values of the Illumina platform, for instance: the model-based background correction method (MBCB) from Ding et al. [12] and Xie et al. [68], the exponential-gamma from Chen et al. [8], the gamma-normal from Plancade et al. [54] and the exponential(gamma)-lognormal from Fajriyah [19].

The study of Posekany et al. [55] by using the Affymetrix and Invitrogen platforms, show that the noises in microarray data are not Gaussian but far more heavy-tailed. In the case of the Illumina platform, Chen et al. [8] show that the noise distribution in the Illumina platform is usually skewed in different degrees.

Therefore, while the intensity values are widely accepted as skewed-distributed, the noise distribution could actually be symmetrically *or* skewed-distributed. Note that in this thesis, noise and intensity mean the negative control probes and the observed probes intensity values respectively.

McDonald and Xu [47] have introduced a distribution tree of generalized beta distributions, which is used to model the income distribution. It is similar in nature to the microarray data where the random variable is a non-negative value. This distribution tree helps us to understand the relationship among the available distributions. Moreover, quite recently, Leemis and McQueston [44] have explained the relationships among the univariate distributions in statistics. See the distribution tree from McDonald and Xu [47] in Figure 3.1.

We aim to present the true intensity value, the corrected background intensity, for both symmetrically distributed noise and skewed-distributed noise. If the noise is a skewed distribution, the underlying distributions of the proposed convolution

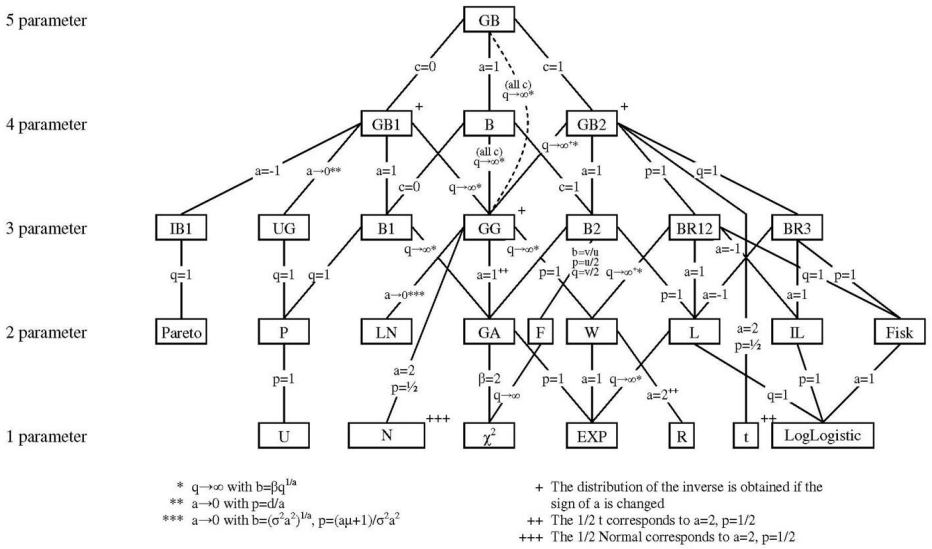


FIGURE 3.1: Distribution tree, [47]

model are the generalized beta distributions, a generalized model of the existing ones. If the noise is a symmetrically distributed, the proposed model is a generalized beta-normal convolution, which is a generalized model of the Plancade et al. model [53, 54].

This section is organized as follows: Section 3.4.2 describes the generalized beta convolution. Section 3.4.3 explains the results of the generalized beta-normal convolution and Section 3.5 provides discussion and remarks.

3.4.2 Generalized beta distribution convolution

3.4.2.1 The joint density function

Under the convolution model of Equation (2.1), where P_i is the observed intensity of regular probes of the i^{th} gene, S_i is the true intensity of the i^{th} gene, with

$$\begin{aligned} S_i &\sim f_1\left(s_i; a_{1,j}, c_{1,j}, g_{1,j}, u_{1,j}, v_{1,j}\right) \\ &= \frac{\left|a_{1,j}\right| s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{g_{1,j}^{a_{1,j}u_{1,j}} B\left(u_{1,j}, v_{1,j}\right)\left(1+c_{1,j}\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}}; \end{aligned} \tag{3.17}$$

$$0 \leq c_{1,j} \leq 1, a_{1,j}, g_{1,j}, u_{1,j} \text{ and } v_{1,j} \text{ positive, } s_i > 0,$$

and B_i is the background noise with

$$\begin{aligned} B_i &\sim f_2\left(b_i; a_{2,j}, c_{2,j}, g_{2,j}, u_{2,j}, v_{2,j}\right) \\ &= \frac{\left|a_{2,j}\right| b_i^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{b_i}{g_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{g_{2,j}^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right)\left(1+c_{2,j}\left(\frac{b_i}{g_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}; \\ &\quad 0 \leq c_{2,j} \leq 1, a_{2,j}, g_{2,j}, u_{2,j} \text{ and } v_{2,j} \text{ positive, } b_i > 0. \end{aligned} \tag{3.18}$$

The joint density function of S_i and B_i is

$$\begin{aligned} f_{S_i,B_i}\left(s_i, b_i\right) =& \frac{\left|a_1\right| s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{g_{1,j}^{a_{1,j}u_{1,j}} B\left(u_{1,j}, v_{1,j}\right)\left(1+c_{1,j}\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}} \times \\ & \frac{\left|a_{2,j}\right| b_i^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{b_i}{g_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{g_{2,j}^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right)\left(1+c_{2,j}\left(\frac{b_i}{g_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}. \end{aligned} \tag{3.19}$$

The joint density function of S_i and P_i is

$$\begin{aligned} & f_{S_i,P_i}\left(s_i, p_i\right) \\ &= \frac{\left|a_1\right| s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{g_{1,j}^{a_{1,j}u_{1,j}} B\left(u_{1,j}, v_{1,j}\right)\left(1+c_{1,j}\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}} \times \\ & \frac{\left|a_{2,j}\right| \left(p_i-s_i\right)^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{(p_i-s_i)}{g_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{g_{2,j}^{a_{2,j}u_{2,j}} B\left(u_{2,j}, v_{2,j}\right)\left(1+c_{2,j}\left(\frac{(p_i-s_i)}{g_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}. \end{aligned} \tag{3.20}$$

3.4.2.2 The marginal density function

The marginal density function of P_i is

$$\begin{aligned} f_{P_i}(p_i) &= \int_0^{p_i} f_{S_i,P_i}(s_i,p_i)\,ds_i \\ &= K\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left\{\frac{(-1)^{l+w+n+r}\left(1-c_{1,j}\right)^{l}\left(1-c_{2,j}\right)^{w}c_{1,j}^{n}c_{2,j}^{r}}{g_{1,j}^{a_{1,j}(l+n)}g_{2,j}^{a_{2,j}(w+r)}}\times\right. \\ &\binom{v_{1,j}-1}{l}\binom{v_{2,j}-1}{w}\binom{u_1+v_1+n-1}{n}\binom{u_{2,j}+v_{2,j}+r-1}{r}\times \\ &\left.\int_0^{p_i} s_i^{a_{1,j}(u_{1,j}+l+n)-1}\left(p_i-s_i\right)^{a_{2,j}\left(u_{2,j}+w+r\right)-1}ds_i\right\}. \end{aligned} \tag{3.21}$$

Let $\frac{s_i}{p_i}=z_i$, therefore, the equation (3.21) becomes

$$K_1 p_i^{a_{1,j}u_1+a_{2,j}u_{2,j}-1}C_{5,j}, \tag{3.22}$$

where

$$K_1 = \frac{|a_{1,j}||a_{2,j}|}{g_{1,j}^{a_{1,j}u_{1,j}}g_{2,j}^{a_{2,j}u_{2,j}}B(u_{1,j},v_{1,j})B(u_{2,j},v_{2,j})}$$

$$\begin{aligned} C_{5,j} = &\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left\{\frac{(-1)^{l+w+n+r}\left(1-c_{1,j}\right)^{l}\left(1-c_{2,j}\right)^{w}c_{1,j}^{n}c_{2,j}^{r}\binom{v_{1,j}-1}{l}}{g_{1,j}^{a_{1,j}(l+n)}g_{2,j}^{a_{2,j}(w+r)}}\times\right. \\ &\binom{v_{2,j}-1}{w}\binom{u_{1,j}+v_{1,j}+n-1}{n}\binom{u_{2,j}+v_{2,j}+r-1}{r}p_i^{a_{1,j}(l+n)+a_{2,j}(w+r)}\times \\ &\left.\mathrm{B}\left(a_{1,j}\left(u_{1,j}+l+n\right)-1,a_{2,j}\left(u_{2,j}+w+r\right)-1\right)\right\}. \end{aligned}$$

3.4.2.3 The conditional density function

The conditional density function of S_i where it is known that $P_i=p_i$ is

$$f_{S_i|P_i}\left(s_i \mid p_i\right) = \frac{f_{S_i,P_i}(s_i,p_i)}{f_{P_i}(p_i)}$$

$$= \frac{s_i^{a_{1,j}u_{1,j}-1}\left(1-\left(1-c_{1,j}\right)\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{v_{1,j}-1}}{p_i^{a_{1,j}u_{1,j}+a_{2,j}u_{2,j}-1}C_{5,j}\left(1+c_{1,j}\left(\frac{s_i}{g_{1,j}}\right)^{a_{1,j}}\right)^{u_{1,j}+v_{1,j}}}\times$$
$$\frac{(p_i-s_i)^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{(p_i-s_i)}{g_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{\left(1+c_{2,j}\left(\frac{(p_i-s_i)}{g_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}. \tag{3.23}$$

3.4.2.4 The corrected background intensity

The corrected background intensity under this generalized beta convolution is

$$E\left(S_i \mid P_i = p_i\right) = \int_0^{p_i} s_i f\left(s_i \mid p_i\right) ds_i = p_i \frac{C_{6,j}}{C_{5,j}}, \tag{3.24}$$

where

$$C_{6,j} = \sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{n=0}^{\infty}\sum_{r=0}^{\infty}\left\{\frac{(-1)^{l+w+n+r}\left(1-c_{1,j}\right)^{l}\left(1-c_{2,j}\right)^{w}c_{1,j}^{n}c_{2,j}^{r}\binom{v_{1,j}-1}{l}}{g_{1,j}^{a_{1,j}(l+n)}g_{2,j}^{a_{2,j}(w+r)}}\times\right.$$
$$\binom{v_{2,j}-1}{w}\binom{u_{1,j}+v_{1,j}+n-1}{n}\binom{u_{2,j}+v_{2,j}+r-1}{r}p_i^{a_{1,j}(l+n)+a_{2,j}(w+r)}\times$$
$$\left.\mathrm{B}\left(a_{1,j}\left(u_{1,j}+l+n\right),a_{2,j}\left(u_{2,j}+w+r\right)-1\right)\right\}.$$

3.4.2.5 The likelihood function

The likelihood function ($\mathbf{L}$) to estimate $a_{1,j}, c_{1,j}, g_{1,j}, u_{1,j}, v_{1,j}, a_{2,j}, c_{2,j}, g_{2,j}, u_{2,j}$, and $v_{2,j}$ is

$$= \prod_{i=1}^{I}\frac{\left|a_{1,j}\right|\left|a_{2,j}\right|p_i^{a_{1,j}u_{1,j}+a_{2,j}u_{2,j}-1}C_{5,j}}{g_{1,j}^{a_{1,j}u_{1,j}}g_{2,j}^{a_{2,j}u_{2,j}}\mathrm{B}\left(u_{1,j},v_{1,j}\right)B\left(u_{2,j},v_{2,j}\right)}\times$$
$$\prod_{m=1}^{M}\frac{\left|a_{2,j}\right|b_{0m}^{a_{2,j}u_{2,j}-1}\left(1-\left(1-c_{2,j}\right)\left(\frac{b_{0m}}{g_{2,j}}\right)^{a_{2,j}}\right)^{v_{2,j}-1}}{g_2^{a_{2,j}u_{2,j}}B\left(u_{2,j},v_{2,j}\right)\left(1+c_{2,j}\left(\frac{b_{0m}}{g_{2,j}}\right)^{a_{2,j}}\right)^{u_{2,j}+v_{2,j}}}. \tag{3.25}$$

The log-likelihood function l is

$$
\begin{aligned}
&= \sum_{i=1}^{I} \Big\{ \ln\left(\left|a_{1,j}\right|\right) + \ln\left(\left|a_{2,j}\right|\right) + \left(a_{1,j}u_{1,j} + a_{2,j}u_{2,j} - 1\right)\ln\left(p_i\right) \\
&\quad + \ln\left(C_{5,j}\right) - \left(a_{1,j}u_{1,j}\right)\ln\left(g_{1,j}\right) - \left(a_{2,j}u_{2,j}\right)\ln\left(g_{2,j}\right) - \ln\Big(\mathrm{B}\left(u_{1,j}, v_{1,j}\right)\Big) \\
&\quad - \ln\Big(\mathrm{B}\left(u_{2,j}, v_{2,j}\right)\Big) \Big\} + \sum_{m=1}^{M} \Big\{ \ln\left(\left|a_{2,j}\right|\right) + \left(a_{2,j}u_{2,j} - 1\right)\ln\left(b_{0m}\right) \\
&\quad + \left(v_{2,j} - 1\right)\ln\left(\left(1 - \left(1 - c_{2,j}\right)\left(\frac{b_{0m}}{g_{2,j}}\right)^{a_{2,j}}\right)\right) - \left(a_{2,j}u_{2,j}\right)\ln\left(g_{2,j}\right) \\
&\quad - \ln\Big(\mathrm{B}\left(u_{2,j}, v_{2,j}\right)\Big) - \left(u_{2,j} + v_{2,j}\right)\ln\left(\left(1 + c_{2,j}\left(\frac{b_{0m}}{g_{2,j}}\right)^{a_{2,j}}\right)\right) \Big\}. \qquad (3.26)
\end{aligned}
$$

3.4.3 Generalized beta-normal convolution

Although Figure 3.1 covers normal distribution, we can not derive the formula of the true intensity value when the noise is normal from Equation (2.1). The normal distribution in Figure 3.1 is the normal distribution with one parameter. Therefore, in this section, we derive the formula to compute the corrected background intensity when the noise is symmetrically distributed, i.e. a normal distribution.

3.4.3.1 The joint density function

Under the convolution model in Equation (2.1), where P_i is the observed intensity of the regular i^{th} gene, S_i is the true intensity of the i^{th}, with

$$
\begin{aligned}
S_i &\sim f_1\left(s_i; a_j, c_j, g_j, u_j, v_j\right) \\
&= \frac{\left|a_j\right| s_i^{a_j u_j - 1}\left(1 - \left(1 - c_j\right)\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{v_j - 1}}{g_j^{a_j u_j}\mathrm{B}\left(u_j, v_j\right)\left(1 + c_j\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{u_j + v_j}}; \qquad (3.27) \\
&\quad 0 \le c_j \le 1; a_j, g_j, u_j, v_j, s_i > 0,
\end{aligned}
$$

and B_i is the background noise with

$$B_i \sim f_2\left(b_i;\mu_j,\sigma_j^2\right)=\frac{e^{-\frac{1}{2\sigma_j^2}\left(b_i-\mu_j\right)^2}}{\sqrt{2\pi}\sigma_j};\mu_j\in\mathbb{R},\sigma_j^2>0,b_i>0. \tag{3.28}$$

The joint density function of S_i and B_m is

$$f_{S_i,B_i}\left(s_i,b_i\right)=\frac{\left|a_j\right|s_i^{a_ju_j-1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{v_j-1}}{g_j^{a_ju_j}\mathrm{B}\left(u_j,v_j\right)\left(1+c_j\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{u_j+v_j}}\frac{e^{-\frac{1}{2\sigma_j^2}\left(b_i-\mu_j\right)^2}}{\sqrt{2\pi}\sigma_j}, \tag{3.29}$$

and the joint density function of S_i and P_i is

$$f_{S_i,P_i}\left(s_i,p_i\right)=\frac{\left|a_j\right|s_i^{a_ju_j-1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{v_j-1}}{g_j^{a_ju_j}\mathrm{B}\left(u_j,v_j\right)\left(1+c_j\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{u_j+v_j}}\frac{e^{-\frac{\left(p_i-s_i-\mu_j\right)^2}{2\sigma_j^2}}}{\sqrt{2\pi}\sigma_j}. \tag{3.30}$$

3.4.3.2 The marginal density function

The marginal density function of P_i is

$$f_{P_i}\left(p_i\right)=\frac{\left|a_j\right|}{g_j^{a_ju_j}\mathrm{B}\left(u,v\right)\sqrt{2\pi}\sigma_j}\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\left\{\frac{(-1)^{l+m}\left(1-c_j\right)^l c_j^w\binom{v_j-1}{l}}{g_j^{a_j(l+w)}}\times\right.$$
$$\left.\binom{u_j+v_j+w-1}{w}\int_0^{p_i}s_i^{a_j(u_j+l+w)-1}e^{-\frac{(s_i-p_i-\mu_j)^2}{2\sigma_j^2}}ds_i\right\}. \tag{3.31}$$

Let $\frac{\left(s_i-\left(p_i-\mu_j\right)\right)}{\sigma_j}=z_i$ and the equation (3.31) becomes

$$
\begin{aligned}
=&\frac{\left|a_j\right|}{g_j^{a_j u_j}\mathrm{B}\left(u_j,v_j\right)\sqrt{2\pi}}\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{n=0}^{\infty}\left\{\frac{(-1)^{l+w}\left(1-c_j\right)^l c_j^w\binom{v_j-1}{l}}{g_j^{a_j(l+w)}\left(p_i-\mu_j\right)^n}\times\right.\\
&\binom{u_j+v_j+w-1}{w}\binom{a_j\left(u_j+l+w\right)-1}{n}\left(p_i-\mu_j\right)^{a_j\left(u_j+l+w\right)-1}\sigma_j^n\times\\
&\left.\int_{-\frac{\left(p_i-\mu_j\right)}{\sigma_j}}^{\frac{\mu_j}{\sigma_j}} z_i^n e^{-\frac{z_i^2}{2}}dz_i\right\}. \qquad (3.32)
\end{aligned}
$$

Let $\frac{z_i^2}{2}=x_i$, then the equation (3.32) becomes

$$=K_2 C_{7,j}. \qquad (3.33)$$

where

$$K_2=\frac{\left|a_j\right|p_i^{a_j u_j-1}}{2\sqrt{\pi}g_j^{a_j u_j}\mathrm{B}\left(u_j,v_j\right)},\text{ and}$$

$$
\begin{aligned}
C_{7,j}=&\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{n=0}^{\infty}\left\{\frac{(-1)^{l+w}\left(1-c_j\right)^l c_j^w\binom{v_j-1}{l}\binom{u_j+v_j+w-1}{w}}{g_j^{a_j(l+w)}\left(p_i-\mu_j\right)^n}\times\right.\\
&\binom{a_j\left(u_j+l+w\right)-1}{n}\left(p_i-\mu_j\right)^{a_j(l+w)}\sigma_j^n 2^{\frac{n}{2}}\left(\gamma\left(\frac{n+1}{2},\left(\frac{\mu_j}{\sigma_j}\right)^2\right)-\right.\\
&\left.\left.\gamma\left(\frac{n+1}{2},\left(\frac{\left(p_i-\mu_j\right)}{\sigma_j}\right)^2\right)\right)\right\},\text{ and}
\end{aligned}
$$

$\gamma\left(\blacksquare,\blacksquare\right)$ is the lower incomplete gamma function

3.4.3.3 The conditional density function

The conditional density function of S_i where it is known that $P_i=p_i$ is

$$f_{S_i|P_i}\left(s_i\mid p_i\right)=\frac{\sqrt{2}p_i^{1-a_j u_j}}{C_{7,j}\sigma_j}\frac{s_i^{a_j u_j-1}\left(1-\left(1-c_j\right)\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{v_j-1}e^{-\frac{\left(p_i-s_i-\mu_j\right)^2}{2\sigma_j^2}}}{\left(1+c_j\left(\frac{s_i}{g_j}\right)^{a_j}\right)^{u_j+v_j}} \qquad (3.34)$$

3.4.3.4 The corrected background intensity

The corrected background intensity under this generalized beta convolution is

$$E\left(S_i \mid P_i = p_i\right) = p_i \frac{C_{8,j}}{C_{7,j}}, \tag{3.35}$$

where

$$C_{8,j} = \sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{n=0}^{\infty}\left\{\frac{(-1)^{l+w}\left(1-c_j\right)^l c_j^w \binom{v_j-1}{l}\binom{u_j+v_j+w-1}{w}}{g_j^{a_j(l+w)}\left(p_i-\mu_j\right)^n}\times \right.$$
$$\binom{a_j\left(u_j+l+w\right)}{n}\left(p_i-\mu_j\right)^{a_j(l+w)}\sigma_j^n 2^{\frac{n}{2}}\left(\gamma\left(\frac{n+1}{2},\left(\frac{\mu_j}{\sigma_j}\right)^2\right)-\right.$$
$$\left.\left.\gamma\left(\frac{n+1}{2},\left(\frac{\left(p_i-\mu_j\right)}{\sigma_j}\right)^2\right)\right)\right\}, \text{and}$$

$\gamma\left(\boldsymbol{.},\boldsymbol{.}\right)$ is the lower incomplete gamma function.

3.4.3.5 The likelihood function

The likelihood function ($\mathbf{L}$) to estimate $a_j, c_j, g_j, u_j, v_j, \mu_j$ and σ_j^2 is

$$= \prod_{i=1}^{I}\frac{\left|a_j\right| p_i^{a_j u_j - 1} C_{7,j}}{2\sqrt{\pi} g_j^{a_j u_j}\mathrm{B}\left(u_j, v_j\right)}\prod_{m=1}^{M}\frac{\exp\left\{-\frac{1}{2\sigma_j^2}\left(b_{0m}-\mu_j\right)^2\right\}}{\sqrt{2\pi}\sigma_j}. \tag{3.36}$$

The log-likelihood function l is

$$= \sum_{i=1}^{I}\left\{\ln\left(\left|a_j\right|\right) + \left(a_j u_j - 1\right)\ln\left(p_i\right) + \ln(C_{7,j}) - \ln(2) - \frac{1}{2}\ln(\pi) - a_j u_j \ln\left(g_j\right) - \right.$$
$$\left.\ln\left(\mathrm{B}\left(u_j, v_j\right)\right)\right\} + \sum_{m=1}^{M}\left\{-\frac{\left(b_{0m}-\mu_j\right)^2}{2\sigma_j^2} - \frac{1}{2}\left(\ln(2)+\ln(\pi)\right) - \ln(\sigma_j)\right\}. \tag{3.37}$$

3.5 Discussion and remarks

We have studied the additive models of the background correction for BeadArrays and proposed the generalized model where the true intensity and the noise are

assumed to be skewed distributions and where the true intensity is a skewed but the noise is symmetrically distribution. In this thesis we have shown the corrected background intensity value of the proposed models.

This proposed model is a generalization of the available convolution models found in [5], [40–42], [8], [19], [40–42], [53, 54] and [68]. The generalization comes from the property of the tree-generalized beta distribution [47] and is explained in [46] and [47]. The parameters of the generalized beta distribution are a, g, c, u and v. The gamma, exponential and lognormal distributions are special cases of the generalized beta distribution.

The gamma distribution is the generalized beta distribution when $c = 1, v \to \infty, g = \beta v^{\frac{1}{a}}$ and $a = 1$; the exponential distribution is the generalized beta distribution when $c = 1, v \to \infty, g = \beta v^{\frac{1}{a}}$ and $a = 1, p = 1$; and the lognormal distribution is the generalized beta distribution when $c = 1, v \to \infty, g = \beta v^{\frac{1}{a}}$ and $\beta = (\sigma^2 a^2)^{\frac{1}{2}}, u = \frac{(a\mu+1)}{\sigma^2 a^2}$ and $a \to 0$.

There are some aspects to be considered while implementing these models:

1. parameters estimation

   In parameters estimation, there are a few methods that have been suggested by researchers. Mc Donald and Xu [47] used and suggested the method of maximum likelihood (also used by Fajriyah [15–17]), the method of moments and the maximum product spacing estimation.

   When $c = 1$, the generalized beta distribution is a generalized beta of the second kind. Graf and Nedyalkova [30] and Graf et al. [31] have observed that the pseudo maximum likelihood (Huber [35], Freedman [25] and Pfeffermann et al. [52]), the nonlinear least squares on the quantile function (Dagum [11]), and the nonlinear fit for indicator can be implemented to estimate the parameters of the generalized beta of the second kind. The available VGAM *package* in R is one of the ways to estimate the parameters of this distribution.

   The existing convolution models use various methods :

   (a) the *ad-hoc* method which is implemented by the RMA method, more details can be found in [40–42], [48] and [68]

   (b) Markov chain Monte Carlo simulations, more details can be found in [12]

(c) Maximum likelihood, nonparametrics and the method of moments, more details can be found in [8], [19] and [68]

(d) Plug-in method, more details can be found in [19]

(e) Fast Fourier transform, more details can be found in [53, 54]

In general, we first need to provide the initial parameters to optimize the log-likelihood function in Equations (3.26) and (3.37). The initial parameters of the noise are easily provided since the benchmarking data set of the negative control probes is available publicly. The initial parameters of the true intensity can be estimated from the observed intensity data substracted by the mean (or median) of the negative control probes intensity.

Secondly, once the initial parameters are available, they can then be used to optimize the likelihood function by implementing the optimization method. There are some packages in R which can be used to compute the parameters of the model, for example the optim or optimx *packages*. These parameters are then used to compute the corrected background intensity based on the formula of the choosen model. Remember that the background correction is implemented for each array.

2. the corrected background intensity computation
   The corrected background intensity computation includes computations of the infinite summations: $C_{5,j}, C_{6,j}, C_{7,j}$ and $C_{8,j}$. In the author's experience (in [19]) these infinite summations become close to being constant after certain terms. As a consequence, the ratios of $\frac{C_{6,j}}{C_{5,j}}$ and $\frac{C_{8,j}}{C_{7,j}}$ are able to be computed. Therefore, the difficulty in computing the summations used to compute the corrected background intensity can be eliminated. A sophisticated programming skill in R, C, *Python* and the parallelisation, could be a great help to speed up the computation.

3. the benchmarking data set
   During the implementation of this generalized estimator, Illumina users need to be aware of the availability of the Illumina spike-in data set. Once the model is fitted into this data set, the model can then be used to adjust the intensity value.

   Apart from the benchmarking criteria for the Affymetrix GeneChips, in the author's knowledge, the benchmarking criteria for the Illumina BeadArrays have not yet been formalized. Some researchers, e.g. [8], [53, 54], [57] and [68]

have developed the criteria to assess which background correction methods perform better than the others for the Illumina BeadArrays. These criteria together with the criteria in the Affycomp *package* ([10] and [39]) can be used as the benchmarking criteria for the Illumina BeadArrays. These have been implemented by Fajriyah [19]. The method which has been used by Shi et al. [58] could also be used to assess the performance of background correction methods.

4. the negative control data set
   It is possible that the negative control probes data set is unavailable. In this case, we can adapt the proposed model to the convolution model for the background correction without the negative control probes intensities, as in the RMA model.

Considering these statements, clearly the application of this generalized model towards other platforms is possible.

3.6 Summary of the chapter

In this chapter, the proposed convolution model for the background correction has been described, for both of symmetric and asymmetric noise distribution. The next chapter will compare the performance of the existing and the proposed models (except the generalized one) based in the Illumina benchmarking and in the public data sets.

Chapter 4

Performance comparison

4.1 Benchmarking

We compare all convolution models: Irizarry et al. [40–42] and Bolstad et al. [5]: RMA (Exponential-Normal), Plancade et al. [53, 54]: Gamma-Normal, Chen et al. [8]: Exponential-Gamma, Xie et al. [68]: Exponential-Normal adjusted for Illumina BeadArrays with maximum likelihood estimation (MLE) for the parameters, Bayesian approach and the moment method, and the proposed models: exponential-lognormal and gamma-lognormal.

We will call the methods above, respectively, as follows: ENr, GN, EG, ENm, ENmc, ENn, ELNn, ELNm, ELNp, GLNn, GLNm, and GLNp. We use the MBCB *package* ([1] and [68]) to adjust the intensity values of models ENr, ENm, ENmc and ENn. Except that, the GN uses the NormalGamma *package* [53].

Table 4.1 shows that the GLNn reproduces the Illumina concentration better than others. The ENr performs most comparably to the GLNn. Note that the computation of Kullback-Leibler is implemented in each array j, based on the nominal concentrations (O) in Table 4.1 and observed intensities (P) in Table 4.2, and the value in each table is the median Kullback-Leibler of $J = 42$ arrays.

The Kullback-Leibler coefficient of each array is computed as $\text{K-L}_j = \sum_{i=1}^{I} X_{ij} \log\left(\frac{X_{ij}}{S_{ij}}\right)$ for arbitrary positive sequences $(X_{1j}, ..., X_{Ij}), (S_{1j}, ..., S_{Ij})$ when it can be negative if $S_{ij} > X_{ij}$ for all or for most i. This is a sign that the S is overestimating X, where X could be O (Table 4.1) or P (Table 4.2). In this case, we should not use

the model. Therefore we exclude the GLNp model from further comparisons. The behavior of GLNp is different from other models, also shown in the supplemental plots.

TABLE 4.1: Reproducibility of each method relating to the Illumina spike-in concentration

Model	RMSE	K-L
ENr	1.346	51,310
ENn	1.407	41,010
ENm	1.483	23,170
ENmc	1.483	23,170
EG	1.470	20,660
GN	1.521	58,480
ELNn	1.411	41,200
ELNm	1.489	21,280
ELNp	1.423	37,800
GLNn	**1.323**	**4,333**
GLNm	1.510	29,630
GLNp	10.700	-115,400

TABLE 4.2: Reproducibility of each method relating to the Illumina spike-in based on the experiment data

Model	RMSE	K-L
ENr	7.251	1,141,000
ENn	7.127	1,062,000
ENm	6.927	926,500
ENmc	6.927	926,200
EG	6.919	907,900
GN	7.100	1,183,000
ELNn	7.124	1,062,000
ELNm	6.904	911,600
ELNp	7.092	1,035,000
GLNn	**6.825**	**793,400**
GLNm	6.937	968,400

Table 4.2 shows how each method reproduces the data from the experiment. We see that GLNn reproduces data better than others, based on the RMSE, and the Kullback-Leibler coefficient.

Tables 4.1 and 4.2 provide insight on how the performance comparison among the models can be conducted further.

In the first part, we compute the adopted Affycomp benchmarking criteria, based on the corrected background data and their log transformation.

Secondly, in the simulation, the MSE_{bc} and the L_1 error is computed based on the log transformation of the experiment and the nominal concentration data.

The log transformation that we use in this paper, respectively, for the benchmarking and the FFPE data sets are as follows:

$$y = \log_2(x + \sqrt{(x^2+1)}) \quad \text{and} \quad y = \log_2(x + 1 + \sqrt{(x^2+1)}) \tag{4.1}$$

where x is the nominal concentration (O) or the observed intensity value (P).

4.1.1 Non-simulation

In Table 4.3 it is shown that the ENr provides the smallest variation and IQR. On the other hand, the GLNn model provides the smallest 99.9% percentiles of log fold change for the non spike-in between replicates. The largest variation, IQR, and 99.9% percentiles, respectively are the GLNm, the ELNm and the GN.

TABLE 4.3: Median SD, IQR and 99.9% percentiles of log fold-change for non spike-in between replicates for each model.

Model	Median SD	IQR	99.90%
ENr	**0.027**	**0.062**	0.415
ENn	0.043	0.089	0.441
ENm	0.069	0.139	0.486
ENmc	0.069	0.139	0.486
EG	0.065	0.134	0.477
GN	0.051	0.098	0.520
ELNn	0.045	0.093	0.442
ELNm	0.071	0.145	0.489
ELNp	0.049	0.100	0.449
GLNn	0.038	0.075	**0.398**
GLNm	0.076	0.080	0.507

In Table 4.4 it is shown that, in general, all methods perform similarly to each other. The GLNn models have the highest signal detect R^2. The GN model has the highest R^2 at low concentration but has the lowest R^2 at high concentration. This means that the GN model works better at low concentration. On the other hand, the ENr proves to work better at medium and high concentrations, which is followed closely by the GLNn model.

If we divide the concentrations into two categories, where high concentration means that the nominal concentration is at least 3pM and low concentration means that

TABLE 4.4: The signal detect R^2 by regressing the Nominal and observed value for each model for the Illumina spike-in.

Model	Signal detect R^2	Low.R^2	Med.R^2	High.R^2
ENr	0.959	0.618	**0.698**	**0.559**
ENn	0.958	0.622	0.695	0.557
ENm	0.957	0.635	0.695	0.558
ENmc	0.957	0.635	0.695	0.558
EG	0.957	0.633	0.695	0.558
GN	0.956	**0.650**	0.697	0.555
ELNn	0.958	0.624	0.695	0.557
ELNm	0.957	0.636	0.694	0.558
ELNp	0.958	0.627	0.695	0.557
GLNn	**0.960**	0.609	0.696	0.558
GLNm	0.956	0.637	0.694	0.558

the nominal concentration is at most 1pM, the GLNn model has the highest R^2 (the data is not shown here). It means, in general and at high concentrations, the GLNn offers a better fit than other models.

As in Table 4.4, Table 4.5 shows that all models have similar performance, although the GLNn model has the highest R^2 of nominal concentration against observed log-fold-change.

TABLE 4.5: The R^2 observed log-fold-change against nominal log-fold-changes for the spike in genes.

Model	Obs-intended-fc.R^2	Obs-(low) int-fc.R^2
ENr	0.976	0.989
ENn	0.974	0.990
ENm	0.972	0.985
ENmc	0.972	0.985
EG	0.972	0.986
GN	0.970	0.987
ELNn	0.974	0.990
ELNm	0.972	0.985
ELNp	0.973	0.990
GLNn	**0.978**	**0.991**
GLNm	0.971	0.984

Table 4.6 provides the results from the computation of the AUC value. The table shows that all models have better accuracy at medium concentrations than at low and high concentrations. The ENr performs very poorly at low concentrations and the GLNm performs best. At high concentrations, the ENr performs the best and is followed by the GLNn. In general, the highest AUC is achieved by all models with the MLE parameter estimation method: GLNm, ELNm, and ENm.

TABLE 4.6: The AUC value for each model.

Model	Low concentration AUC	Medium concentrations AUC	High concentration AUC	Average AUC	All
ENr	0.450	0.987	**0.785**	0.585	0.886
ENn	0.518	0.987	0.764	0.631	0.899
ENm	0.573	0.987	0.741	0.667	0.911
ENmc	0.573	0.987	0.741	0.667	0.911
EG	0.567	0.987	0.746	0.664	0.910
GN	0.552	0.987	0.723	0.651	0.904
ELNn	0.524	0.987	0.763	0.635	0.900
ELNm	0.574	0.987	0.741	0.668	0.912
ELNp	0.534	0.987	0.761	0.642	0.902
GLNn	0.498	**0.987**	0.784	0.619	0.896
GLNm	**0.579**	0.987	0.730	**0.671**	**0.913**

The computation, which is based on the 12 and all arrays, provides the results where all models have an AUC greater than 0.9. According to Zhu et al. [71], an AUC between 0.9 and 1.0 is classified as excellent in measuring the accuracy. Therefore, based on Table 4.7, we can identify thos models which accurately predict the gene expression.

In the Appendix Sections A, B and C, we prsent graphics supplementation. The MA plots A.1, A.2 and A.3 show that all models perform similarly, except the GLNp model. In variance across replicates (Figures B.1 and B.2), the GLNn model performs better than other models at low and medium concentrations. At high concentrations the EGm and the GN models perform not at best. The computation (not shown) also results in the GN model producing more differentially expressed non spike-in genes.

A slight over-estimation is shown in figures C.1 and C.2, where all models are above the ideal line at low and medium concentrations, particularly the GN and GLNm models, and then gradually go under the ideal line at high concentrations.

4.1.2 Simulation

We run simulations ($N = 100$) to assess the performance of each model. The bias of the background correction is assessed by the MSE_{bc}, and the bias of the parameterization is assessed by L_1 error. For further details on the simulations see Appendix F.1.

in Table 4.7 we can see that simulation results of the EG model are not available, because the MBCB *package* did not work in the log transformation that we have

TABLE 4.7: The simulation results on spike-in data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	0.045		0.664	46.580	11.440
ENn	0.049		0.625	41.610	2.806
ENm	0.038		0.610	58.920	2.040
ENmc	0.036		0.610	62.770	2.039
GN	**0.030**	**0.000**	**0.007**	0.013	**0.015**
ELNn	0.048		0.009	0.000	0.018
ELNm	0.039		0.840	0.000	0.018
ELNp	0.061		0.472	0.000	0.018
GLNn	0.216	0.052	0.055	0.000	0.018
GLNm	84.370	38.860	0.851	**0.000**	0.017

chosen. The GN model performs best, by providing the smallest bias for the background correction and the parameters. Similar performance is achieved by the ELN, particularly ELNn. The GLNn does not have an optimal performance on the MSE_{bc}, but we still can consider its performance to be good, considering that the bias of the parameters is similar to the other proposed models and GN.

One of the proposed models, GLNm has the highest bias on the MSE_{bc} and the parameter α. In our view this happens because we use an approximation in estimating the true intensity value. The EN models (ENr, ENm, ENn and ENmc) have considerably better performance at MSE_{bc}, but are not good at the parametrization. The bias on the parametrization of the noise is higher than in other models.

4.2 The public data sets

Based on the results from Section 4.3, we compare performance of these models on some public data sets. We would like to know how well these models perform in real data samples. Here, we choose to use the FFPE data sets from Waldron et al. [66]: the FFPE of tumors from colorectal cancer patients (GSE32651, 1003 samples), breast cancer metastases of the lymph node and autopsy tissues (GSE32490: GSE32489, 120 samples). Each sample has 24,526 bead-type level probes.

Links for the data set are http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32651 and http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32490.

Currently, the FFPE archival samples are available by the millions and are a great source of information in medical studies about some diseases, for example

cancer. This data type is suffering from RNA degradation, which leads to poor performance in array-based studies. However, the Illumina's DASL assays could provide high-quality data from these degraded RNA samples.

Comparing the performance of these background correction models would certainly help researchers to choose the appropriate background correction for their data, particularly if their data is of the FFPE type.

The background correction for the FFPE data set is implemented in three steps:

step 1 Apply the quality control (QC) to the raw FFPE data. This study uses the ffpe *package* in R [65].

step 2 Apply the data transformation ($\log_2((P_{ij} + 1 + \sqrt{(P_{ij}^2 + 1)}))$) to the raw FFPE data after QC and estimate the background correction parameters based on its result. The estimators of true intensity value and the background correction are based on the regular and negative control probes intensity data respectively.

step 3 Compute the true intensity value (the adjusted intensity estimator) based on the background correction parameters in *step 2*.

The results of our computation are presented in Tables 4.8 and 4.9. From these tables, we can see that there are no EG and GN models. Neither of these models can work for these data sets. For some samples in the data set, both models fail to compute the parameters which consequently, that the true intensity value cannot be provided.

We decided to remove the EG and GN models from further comparisons in both FFPE data sets. Hereafter we provide the results of the rest of the models only.

Tables 4.8 and 4.9 consistently show the biases of the noise's parameters in the EN models are higher than the proposed models. For the parameter β, the ELNn has the smallest bias and is followed by the ELNp and the GLNn. With regard to the bias of the background correction, the EN models show the smallest bias in both FFPE data sets.

The proposed model GLNm continues to show the highest bias in the background correction and the parameter α. As previously mentioned, this is a consequence of the approximate computation of the true intensity value, where we make the

TABLE 4.8: Simulation results of the GSE32651 data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	0.058		0.657	297.671	22.609
ENn	0.281		0.572	1.984	2.627
ENm	0.086		0.591	28.049	1.715
ENmc	**0.036**		0.610	62.770	2.039
ELNn	0.275		**0.025**	0.001	0.018
ELNm	0.059		0.826	**0.000**	0.018
ELNp	0.672		0.487	0.001	0.018
GLNn	0.838	**0.340**	0.527	0.001	0.018
GLNm	84.150	71.870	0.887	0.001	**0.018**

TABLE 4.9: Simulation results of the GSE32489 data set.

Model	MSE_{bc}	L_1 error			
		α	β	μ	σ
ENr	**0.093**		0.665	67.170	9.511
ENn	0.863		0.509	0.936	2.039
ENm	0.182		0.558	14.197	1.712
ENmc	0.184		0.556	14.179	1.049
ELNn	1.055		0.857	0.002	0.018
ELNm	0.116		0.781	0.001	0.0178
ELNp	1.247		0.461	0.002	0.018
GLNn	1.348	**0.332**	**0.497**	0.002	**0.018**
GLNm	164.980	22.239	0.805	**0.000**	**0.018**

approximation until $k = 10$. It ispossible to apply different numerical approximations for both the ELN and GLN models.

4.3 The risk ratio comparison

4.3.1 Motivation

Background correction plays an important role in microarray data processing, since some steps in producing microarray data contribute the noise. Since Irizarry et al. [40] introduced the convolution model for the Affymetrix platform single channel, it has also been impelemnted in other platforms, such as two-colours/channels and BeadArrays, where for each platform the adaptive models have been developed.

We have developed the exponential and gamma-lognormal convolution models [19] and compared their performances to the existing models. Here, we study the mean

absolute deviation of the BC of all existing convolution models and compare their risk ratio.

4.3.2 Measuring the risk ratio

The excess risk ratio, of using particular model i where the *true* model j is known, is defined as follows:

$$R(i) = \frac{MAD(\widehat{S}_i)}{MAD(\widehat{S}_j)}, \tag{4.2}$$

where

$$\text{MAD}(\widehat{S}) = \frac{1}{N}\sum_{l=1}^{N}\left(\frac{1}{n_r}\sum_{k=1}^{n_r}\left|\widehat{S}(X_k^l \mid \widehat{\Theta}_l) - S_k^l\right|\right), \tag{4.3}$$

X is the intensity value of regular probes, $\widehat{\Theta}_l$ is the estimated parameters from the simulated array l, and S is the true intensity value.

We compare the excess risk ratio of the existing models through the simulation based on the Illumina benchmarking data set. The simulation is conducted as follows:

1. Select a particular model as the *true* model.
2. Generate the samples: regular and negative control, based on the parameters of this *true* model, for each model.
3. Estimate the parameters and the true intensity value for each model.
4. Compute the MAD value.
5. Compute the risk ratio.

Respectivley, the sample size of the regular probes is 25000(n_r), the negative control probes is 1000 and the simulation is run $N = 100$ times.

4.3.3 Results

Table 4.10 shows that then*GN* model produce a *NaN* value when it is assumed that the true model is ENr. This result is quite surprising, since the parameters are based on the benchmarking data set. By excluding this value from the computation and computing the average of the risk ratio for each model we discover that the GN model produces the lowest risk ratio. It is followed by the ENn, ELNn, ELNp, ENm, ENmc, GLNm, GLNp, and ENr.

This unexpected *NaN* value shows that the GN model needs to be carefully implemented. As described in Section 4.2, this model and the EG model could not be used in the FFPE public data set.

TABLE 4.10: Comparison of the risk ratio for each model

True | Applied	**ENr**	**ENn**	**ENm**	**ENb**	**GN**	**ELNn**	**ELNm**	**ELNp**	**GLNn**	**GLNm**
ENr	1.000	0.638	0.638	0.638	**NaN**	0.824	46.497	0.961	2.796	1.434
ENn	2.871	1.000	1.000	1.000	1.000	1.003	42.762	1.040	1.091	1.098
ENm	3.199	1.0005	1.000	1.000	1.000	1.000	47.832	1.015	1.020	1.060
ENmc	3.204	1.000	1.000	1.000	1.000	1.001	47.347	1.015	1.019	1.059
GN	1.836	1.232	1.903	1.903	1.000	1.237	56.436	1.296	3.294	1.819
ELNn	2.809	0.998	0.997	0.998	0.998	1.000	42.931	1.034	1.086	1.092
ELNm	0.071	0.022	0.0228	0.022	0.022	0.022	1.000	0.022	0.022	0.023
ELNp	4.484	1.000	1.000	1.000	1.000	1.000	38.481	1.000	1.000	1.002
GLNn	0.614	0.410	0.590	0.590	0.341	0.413	16.082	0.438	1.000	1.357
GLNm	3.525	0.922	0.943	0.943	0.923	0.922	41.740	0.927	0.938	1.000

4.4 Discussion

We have compared the performance of all models, based on the benchmarking and public data sets. In the benchmarking data set we adopted the criteria from the Affycomp [10] and for the simulation study we used the criteria which had been used in [68], [8] and [53, 54]. For the public data sets, we only used the criteria for the simulation study.

We have seen in Sections 4.1.1 and 4.1.2 that EN, EG, GN and GLN perform rather similarly. However, the GLNn model provides the highest reproducibility in comparison to other models. From the Affycomp criteria we can provide the following points:

1. the ENr and GLNn provide the lowest variation between replicates and all models using the MLE estimation method have a higher variation than others

2. the GLNn model has the highest signal detect R^2, in general and in high concentration. This means the GLNn model is the best fitted for the gene expression.

3. the GLNn model, based on the MvA plot, produces the least number of genes which should not be expressed but are nevertheless expressed. On the other hand, the GN model provides the largest number of such genes.

4. all models with the MLE estimation method have a higher average AUC value, which means that they provide better accuracy in predicting the gene expression.

5. the ENr and GLNn have the lowest IQR of log fold-change between replicates

6. Points 1 and 2 show that the GLNn and ENr are more accurate and precise in modelling the gene expression and points 3 and 5 show that the specificity and sensitivity of the GLNn and ENr model are better than others.

In the simulation study, the best performance in estimating the signal by measuring its background correction and parametrization errors is achieved by the GN model. It is followed by our proposed ELN models. It has been shown that the GLNn does not perform optimally for the MSE_{bc} criterion, but for the parametrization this model still can be considered good.

In the FFPE public data set, the GN and EG models cannot be implemented. This is in strong contrast to the fact that in the simulation study of the benchmarking data sets, the GN model has the best performance.

The EN models show both the highest bias in parametrization of public data sets and the lowest bias in background correction. Our proposed models, except the GLNm, show the lowest bias in parametrization in both data sets and a moderate bias in background correction.

Based on the results from the benchmarking data and the public data sets, we would make the following suggestions to researchers:

1. if the GN model works properly for the data set at hand (i.e. the estimated signals in all arrays can be computed by this model and the simulation criteria for this data with this model are low) then use the GN model to correct the background.

2. if the GN model fails, then use our proposed models, particularly the GLNn model. The reason for not choosing the ELN models is that the value of the parameter α from the benchmarking data set is less than 1, around 0.2. Therefore, the gamma model is more appropriate to model the true intensity distribution than the exponential model. We believe that the right approximative computation of the GLN models will lead to better performance than the current approximation.

   The ELN models perform better than the original EN models, due to the fact that not only the regular probes, but also the control probes are skew-distributed [8]. Therefore, these models could be an alternative to the GLN, when the GN model does not work.

3. With regard to the computation time, for the benchmarking data set the EN models work faster than the others. They are followed by the ELNp, ELNn, and EGm. The GLNn and the ENmc are the third fastest, then come the GN and the ELNm, which are followed by the GLNm, which is the slowest.

In measuring the risk based on the simulation study, we conclude that:

1. The risk ratio of the Exponential-Normal models, except ENr is always lower than 1.

2. The Gamma-Normal model behaves unpredictably in regard to measuring the risk ratio. There is a situation where this model can not be implemented, because it produces a *NaN* value. This behaviour is similar to the previous result of Fajriyah [18].

3. The proposed models provide the best performance by showing a consistently moderate risk ratio, particulalry the ELNn and ELNp models where the risk ratio is lower than 1.

4. In general the choice of background correction method is up to the user, who must compromise between low risk ratio, low bias in parametrization, and low bias in background correction.

4.5 Summary of the chapter

In this chapter we have described the performance comparison of all existing and proposed models in the benchmarking and public data sets. We also compute and compare the risk of all models. Our study shows that the proposed models, except the GLNp, are moderately good in both benchmarking and simulation, and are work well with the FFPE public data sets, [19].

According to analysis, in general the GN provides the smallest risk, but it has to be carefully implemented, because there is a possibility that the risk can not be computed. The proposed models (except the ELNm) and the EN models (except the ENr) provide a moderate risk [18].

The next chapter will introduce the proposed test to determine the differentially expressed genes under two conditions, which proposes an alternative test to the two independent samples t test.

Chapter 5

Cross variance and its application

In microarray data analysis, sometimes we deal with raw data which need to be adjusted to clean them of noise. This can be achieved by pre-processing. Once the raw microarray data have been pre-processed, they are ready for further analysis, such as the determination genes which are differentially expressed.

Typical examples of differentially expressed genes are found when investigating genes related to the effect of drugs on cancer tissue versus normal tissue, or on two different types of tissues, or during two different stages of disease. Such investigations require parametric and non-parametric statistical tests for comparing two independent samples.

One of the parametric tests widely used in bioinformatics research is the t test. This test requires representative samples be taken to measure variability, which microarray experiments sometimes fail to provide. The results are, however, inconclusive, because the sample size is not large enough to accurately measure the difference.

In the study of Bryant et al. [7], it is concluded that most of the genes expressions from microarray experiments have both low technical and low biological variations, although the technical variation is higher than the biological. Furthermore, prior to statistical analysis, after the pre-processing, many researchers applied filtering, as described for example in Bourgon et al. [6], Hackstadt and Hess [32], Iterson et al [63], Smyth [60] and von Heydebreck et al. [64].

In the filtering which is applied to the Affymetrix platform, the low variation genes expressions are filtered out (Bourgon et al. [6], Hackstadt and Hess [32], Iterson

et al [63] and von Heydebreck et al. [64]), which is not a suitable choice for those expressions which come from other platforms, such as Illumina.

This chapter describes a new approach, based on variation, to determining the differentially expressed genes under two conditions, leading to the problem of testing for the equality of the mean of two populations. It is constructed based on the cross variance concept.

5.1 Introduction of cross variance

Definition 5.1. Suppose we have two independent samples, X_i and Y_j; $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Their sample mean and variance are denoted by $\overline{X}, \overline{Y}$ and V_x, V_y. Let

$$V_x^* = \frac{\sum_{i=1}^{m}(X_i - \overline{Y})^2}{m-1} \quad \text{and} \quad V_y^* = \frac{\sum_{i=1}^{n}(Y_i - \overline{X})^2}{n-1},$$

be the cross variance for each sample X and Y respectively. The cross variance of samples $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathbf{T} = \frac{V_x^a + V_y^a}{2}, \tag{5.1}$$

where $V_x^a = \frac{V_x}{V_x^*}$, $V_y^a = \frac{V_y}{V_y^*}$.

Clearly

$$V_x^* = \frac{\sum_{i=1}^{m}(X_i - \overline{Y})^2}{m-1} = V_x + \frac{m(\overline{Y} - \overline{X})^2}{m-1},$$

and

$$V_y^* = \frac{\sum_{i=1}^{n}(Y_i - \overline{X})^2}{n-1} = V_y + \frac{n(\overline{Y} - \overline{X})^2}{n-1}.$$

Thus $T = \frac{V_x^a + V_y^a}{2}$ can be re-written as

$$\begin{aligned} T &= \frac{1}{2}\left[\frac{V_x}{V_x + \frac{m(\overline{Y}-\overline{X})^2}{m-1}} + \frac{V_y}{V_y + \frac{n(\overline{Y}-\overline{X})^2}{n-1}}\right] \\ &= \frac{V_x}{2V_x + 2\frac{m(\overline{Y}-\overline{X})^2}{m-1}} + \frac{V_y}{2V_y + 2\frac{n(\overline{Y}-\overline{X})^2}{n-1}} \end{aligned} \tag{5.2}$$

In what follows, we assume that

1. the sample sizes are equal
2. X_i and Y_i are i.i.d. normally distributed with unknown means and known variances σ_x^2, σ_y^2.

It follows that

$$\frac{(n-1)V_x}{\sigma_x^2} \backsim \chi^2_{(n-1)}, \quad \frac{(n-1)V_y}{\sigma_y^2} \backsim \chi^2_{(n-1)}, \quad \text{and} \quad \frac{n(\overline{Y}-\overline{X})^2}{\sigma_y^2+\sigma_x^2} \backsim \chi^2_{(1)}$$

Therefore Equation (5.2) can be written as follows

$$\begin{aligned} T = &\frac{\frac{(n-1)V_x}{\sigma_x^2}}{2\frac{(n-1)V_x}{\sigma_x^2} + 2\frac{(\sigma_x^2+\sigma_y^2)}{\sigma_x^2}\frac{n}{(\sigma_x^2+\sigma_y^2)}(\overline{Y}-\overline{X})^2} + \\ &\frac{\frac{(n-1)V_y}{\sigma_y^2}}{2\frac{(n-1)V_y}{\sigma_y^2} + 2\frac{(\sigma_x^2+\sigma_y^2)}{\sigma_y^2}\frac{n}{(\sigma_x^2+\sigma_y^2)}(\overline{Y}-\overline{X})^2}, \end{aligned} \tag{5.3}$$

where

$$Z_1 = \frac{\frac{(n-1)V_x}{\sigma_x^2}}{\frac{2(n-1)V_x}{\sigma_x^2} + \frac{2(\sigma_x^2+\sigma_y^2)}{\sigma_x^2}\frac{n}{(\sigma_x^2+\sigma_y^2)}(\overline{Y}-\overline{X})^2} = \frac{U}{2U+2abV},$$

and

$$Z_2 = \frac{\frac{(n-1)V_y}{\sigma_y^2}}{\frac{2(n-1)V_y}{\sigma_y^2} + \frac{2(\sigma_x^2+\sigma_y^2)}{\sigma_y^2}\frac{n}{(\sigma_x^2+\sigma_y^2)}(\overline{Y}-\overline{X})^2} = \frac{S}{2S+2bcV}$$

with

$$U = \frac{(n-1)V_x}{\sigma_x^2}, \quad S = \frac{(n-1)V_y}{\sigma_y^2}, \quad V = \frac{n(\overline{Y}-\overline{X})^2}{(\sigma_x^2+\sigma_y^2)}$$

and

$$a = \frac{1}{\sigma_x^2}, \quad b = \sigma_x^2 + \sigma_y^2, \quad c = \frac{1}{\sigma_y^2}.$$

Hence Equation (5.3) can be written as

$$T = Z_1 + Z_2 = \frac{U}{2U + 2abV} + \frac{S}{2S + 2bcV} \tag{5.4}$$

To compute the distribution of T in Equation (5.5), consider that

1. U, V and S are independent
2. Z_1 and Z_2 are dependent

In this chapter we will describe both considerations, respectively at sections 5.2 and 5.3.

5.2 The first proposed test: an alternative to the t test

Under normality asumption of X and Y then U, V and S are independent, where V is $\chi^2_{(1)}$ distributed and U, S are $\chi^2_{(n-1)}$ distributed. From Equation (2.5), suppose $V = Z_3$ from here it follows that $U = \frac{2abZ_1Z_3}{1-2Z_1}$ and $S = \frac{2bcZ_2Z_3}{1-2Z_2}$. The Jacobian of this transformation is

$$|J| = \begin{vmatrix} \frac{dU}{dZ_1} & \frac{dU}{dZ_2} & \frac{dU}{dZ_3} \\ \frac{dS}{dZ_1} & \frac{dS}{dZ_2} & \frac{dS}{dZ_3} \\ \frac{dV}{dZ_1} & \frac{dV}{dZ_2} & \frac{dV}{dZ_3} \end{vmatrix} = \begin{vmatrix} \frac{2abZ_3}{(1-2Z_1)^2} & 0 & \frac{2abZ_1}{(1-2Z_1)} \\ 0 & \frac{2bcZ_3}{(1-2Z_2)^2} & \frac{2bcZ_2}{(1-2Z_2)} \\ 0 & 0 & 1 \end{vmatrix} = \frac{4ab^2cZ_3^2}{\big((1-2Z_1)(1-2Z_2)\big)^2}. \tag{5.5}$$

The joint probability density function of Z_1, Z_2, Z_3 is

$$\begin{aligned} & f_{Z_1,Z_2,Z_3}(z_1, z_2, z_3) \\ &= f_U\left(u = \frac{2abz_1z_3}{1-2z_1}\right) f_S\left(s = \frac{2bcz_2z_3}{1-2z_2}\right) f_V(v = z_3)|J| \\ &= \frac{\left(4ab^2c\right)^{\frac{n-1}{2}}}{2^{n-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2} \frac{z_1^{\frac{n-1}{2}-1} z_2^{\frac{n-1}{2}-1} z_3^{\frac{2n-1}{2}-1} e^{-z_3\left(\frac{1}{2}+\frac{abz_1}{1-2z_1}+\frac{cbz_2}{1-2z_2}\right)}}{(1-2z_1)^{\frac{n+1}{2}}(1-2z_2)^{\frac{n+1}{2}}}. \end{aligned} \tag{5.6}$$

The joint density function of Z_1, Z_2 is the marginal probability function of Z_1, Z_2 from Equation (5.6) above. It is computed as follows:

$$
\begin{aligned}
& f_{Z_1,Z_2}(z_1,z_2) \\
&= \int_0^\infty f_{Z_1,Z_2,Z_3}(z_1,z_2,z_3)dz_3 \\
&= \frac{\left(4ab^2c\right)^{\frac{n-1}{2}}}{2^{n-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2} \frac{z_1^{\frac{n-1}{2}-1} z_2^{\frac{n-1}{2}-1}}{(1-2z_1)^{\frac{n+1}{2}}(1-2z_2)^{\frac{n+1}{2}}} \int_0^\infty z_3^{\frac{2n-1}{2}-1} e^{-z_3\left(\frac{1}{2}+\frac{abz_1}{1-2z_1}+\frac{bcz_2}{1-2z_2}\right)} dz_3 \\
&= \frac{\left(4ab^2c\right)^{\frac{n-1}{2}} \Gamma\left(n-\frac{1}{2}\right) z_1^{\frac{n-1}{2}-1} z_2^{\frac{n-1}{2}-1}}{2^{n-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2 (1-2z_1)^{\frac{n+1}{2}}(1-2z_2)^{\frac{n+1}{2}} \left(\frac{1}{2}+\frac{abz_1}{(1-2z_1)}+\frac{bcz_2}{(1-2z_2)}\right)^{n-\frac{1}{2}}}.
\end{aligned}
\tag{5.7}
$$

Therefore the cumulative distribution function (cdf) of $T \leq t$ is computed as follows:

$$
\begin{aligned}
& F_T(t) \\
&= P(T \leq t) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{t-z_1} f_{Z_1,Z_2}(z_1,z_2)dz_2dz_1 \\
&= \int_0^t \int_0^{t-z_1} \frac{\left(4ab^2c\right)^{\frac{n-1}{2}} \Gamma\left(n-\frac{1}{2}\right) z_1^{\frac{n-1}{2}-1} z_2^{\frac{n-1}{2}-1} dz_2dz_1}{2^{n-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2 \left((1-2z_1)(1-2z_2)\right)^{\frac{n+1}{2}} \left(\frac{1}{2}+\frac{abz_1}{(1-2z_1)}+\frac{bcz_2}{(1-2z_2)}\right)^{n-\frac{1}{2}}} \\
&= \int_0^t \frac{\left(4ab^2c\right)^{\frac{n-1}{2}} \Gamma\left(n-\frac{1}{2}\right) z_1^{\frac{n-1}{2}-1}}{2^{n-\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2 (1-2z_1)^{\frac{n+1}{2}}} \mathbf{B} dz_1
\end{aligned}
\tag{5.8}
$$

where

$$
\mathbf{B} = \int_0^{t-z_1} \frac{z_2^{\frac{n-1}{2}-1}(1-2z_2)^{\frac{n+1}{2}} dz_2}{\left(\frac{1}{2}+\frac{abz_1}{(1-2z_1)}+\frac{bcz_2}{(1-2z_2)}\right)^{n-\frac{1}{2}}}. \tag{5.9}
$$

To compute the integral Equation (5.9), first we simplify this $\left(\frac{1}{2}+\frac{abz_1}{(1-2z_1)}+\frac{bcz_2}{(1-2z_2)}\right)^{n-\frac{1}{2}}$ as $\left(1+\frac{2(bc-(1-2(1-ab-bc)z_1))}{1-2(1-ab)z_1}z_2\right)\left(\frac{1-2(1-ab)z_1}{2(1-2z_1)}\right)$. Therefore Equation (5.9) becomes

$$= \frac{\big(2(1-2z_1)\big)^{n-\frac{1}{2}}}{\big(1-2(1-ab)z_1\big)^{n-\frac{1}{2}}} \int_0^{t-z_1} \frac{z_2^{\frac{n-1}{2}-1}\,(1-2z_2)^{\frac{n+1}{2}}\,dz_2}{\Big(1+\frac{2(bc-(1-2(1-ab-bc)z_1))}{1-2(1-ab)z_1}z_2\Big)^{n-\frac{1}{2}}}. \tag{5.10}$$

The integral $\int_0^{t-z_1} \frac{z_2^{\frac{n-1}{2}-1}(1-2z_2)^{\frac{n+1}{2}}dz_2}{\left(1+\frac{2(bc-(1-2(1-ab-bc)z_1))}{1-2(1-ab)z_1}z_2\right)^{n-\frac{1}{2}}}$ is written as

$$\int_0^{t-z_1} z_2^{\frac{n-1}{2}-1}\,(1-2z_2)^{\frac{n+1}{2}} \left(1+\frac{2(bc-(1-2(1-ab-bc)z_1))}{1-2(1-ab)z_1}z_2\right)^{-\left(n-\frac{1}{2}\right)} dz_2. \tag{5.11}$$

By considering the binomial expansion then Equation (5.11) can be represented as

$$\begin{aligned}
&= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \left[\frac{2^{k+l}(-1)^{k+l}\binom{\frac{n+1}{2}}{k}\binom{n-\frac{3}{2}+l}{l}(bc-1+2(1-ab-bc)z_1)^l}{\big(1-2(1-ab)z_1\big)^l} \times \right. \\
&\left. \int_0^{t-z_1} z_2^{\frac{n-1}{2}+k+l}dz_2 \right] \\
&= \sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \left[\frac{2^{k+l}(-1)^{k+l}\binom{\frac{n+1}{2}}{k}\binom{n-\frac{3}{2}+l}{l}(bc-1+2(1-ab-bc)z_1)^l}{\big(1-2(1-ab)z_1\big)^l} \times \right. \\
&\left. \frac{(t-z_1)^{\frac{n+1}{2}+k+l}}{\frac{n+1}{2}+k+l} \right].
\end{aligned} \tag{5.12}$$

Therefore,

$$\begin{aligned}
\mathbf{B} =& \frac{\big(2(1-2z_1)\big)^{n-\frac{1}{2}}}{\big(1-2(1-ab)z_1\big)^{n-\frac{1}{2}}} \times \\
& \left[\sum_{k=0}^{\infty}\sum_{l=0}^{\infty} \left[\frac{2^{k+l}(-1)^{k+l}\binom{\frac{n+1}{2}}{k}\binom{n-\frac{3}{2}+l}{l}(bc-1+2(1-ab-bc)z_1)^l}{\big(1-2(1-ab)z_1\big)^l} \times \right. \right. \\
& \left. \left. \frac{(t-z_1)^{\frac{n+1}{2}+k+l}}{\frac{n+1}{2}+k+l} \right] \right],
\end{aligned} \tag{5.13}$$

and

$$F_T(t) = \frac{\left(4ab^2c\right)^{\frac{n-1}{2}}\Gamma\left(n-\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2}\times \left[\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\frac{(-2)^k\binom{\frac{n+1}{2}}{k}\binom{n-\frac{3}{2}+l}{l}(-2(bc-1))^l t^{\frac{n+1}{2}+k+l}}{\frac{n+1}{2}+k+l}\mathbf{G}\right], \tag{5.14}$$

where

$$\mathbf{G} = \int_0^t \left[z_1^{\frac{n-1}{2}-1}\left(1-\frac{z_1}{t}\right)^{\frac{n+1}{2}+k+l}(1-2z_1)^{\frac{n}{2}-1}\left(1-2(1-ab)z_1\right)^{-(n+l-\frac{1}{2})}\times \left(1+\frac{2(1-ab-bc)z_1}{bc-1}\right)^l dz_1\right]. \tag{5.15}$$

Again, by considering the binomial expansion then Equation (5.15) can be written as follows

$$\begin{aligned}\mathbf{G} &= \sum_{w=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\left[\binom{\frac{n}{2}}{w}\binom{l}{p}\binom{n+l+q-\frac{3}{2}}{q}(-2)^w\left(\frac{2(1-ab-bc)}{bc-1}\right)^p\times (-2(1-ab))^q\int_0^t z_1^{\frac{n-1}{2+w+p+q}-1}\left(1-\frac{z_1}{t}\right)^{\frac{n-1}{2}+k+l}dz_1\right]\\ &= \sum_{w=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\left[\binom{\frac{n}{2}}{w}\binom{l}{p}\binom{n+l+q-\frac{3}{2}}{q}(-2)^w\left(\frac{2(1-ab-bc)}{bc-1}\right)^p\times \frac{(-2(1-ab))^q\mathrm{B}\left(\frac{n-1}{2}+w+p+q,\frac{n+1}{2}+k+l\right)}{t^{\frac{n-1}{2}+w+p+q}}\right].\end{aligned} \tag{5.16}$$

Therefore $F_T(t)$ is

$$F_T(t) = \frac{\left(4ab^2c\right)^{\frac{n-1}{2}}\Gamma\left(n-\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2}\times \left[\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\left[\frac{2^k\binom{\frac{n-1}{2}+k}{k}\binom{n-\frac{3}{2}+l}{l}(-2(bc-1))^l t^{\frac{n-1}{2}+k+l}}{\frac{n-1}{2}+k+l}\times\right.\right.$$

$$\frac{\binom{\frac{n}{2}}{w}\binom{l}{p}\binom{n+l+q-\frac{3}{2}}{q}(-2)^w\left(\frac{2(1-ab-bc)}{bc-1}\right)^p(-2(1-ab))^q}{t^{\frac{n-1}{2}+w+p+q}}\times$$
$$B\left(\frac{n-1}{2}+w+p+q,\frac{n+1}{2}+k+l\right)\Bigg]\Bigg]$$
$$=\frac{\left(4ab^2c\right)^{\frac{n-1}{2}}\Gamma\left(n-\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2}\times$$
$$\Bigg[\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\Bigg[\frac{2^k\binom{\frac{n-1}{2}+k}{k}\binom{n-\frac{3}{2}+l}{l}(-2(bc-1))^l t^{k+l-w-p-q}}{\frac{n-1}{2}+k+l}\times$$
$$\binom{\frac{n}{2}}{w}\binom{l}{p}\binom{n+l+q-\frac{3}{2}}{q}(-2)^w\left(\frac{2(1-ab-bc)}{bc-1}\right)^p(-2(1-ab))^q\times$$
$$B\left(\frac{n-1}{2}+w+p+q,\frac{n+1}{2}+k+l\right)\Bigg]\Bigg]. \tag{5.17}$$

Furthermore, from Equation (5.17) it follows that the probability density function (pdf) of T is

$$f_T(t)=\frac{\left(4ab^2c\right)^{\frac{n-1}{2}}\Gamma\left(n-\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)^2}\sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\sum_{w=0}^{\infty}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty}\Bigg[\frac{2^k\binom{\frac{n-1}{2}+k}{k}\binom{n-\frac{3}{2}+l}{l}}{\frac{n-1}{2}+k+l}\times$$
$$(k+l-w-p-q)t^{k+l-w-p-q}(-2(bc-1))^l\binom{\frac{n}{2}}{w}\binom{l}{p}\times$$
$$\binom{n+l+q-\frac{3}{2}}{q}(-2)^w\left(\frac{2(1-ab-bc)}{bc-1}\right)^p(-2(1-ab))^q\times$$
$$B\left(\frac{n-1}{2}+w+p+q,\frac{n+1}{2}+k+l\right)\Bigg] \tag{5.18}$$

The pdf of T can also be computed as follows:

$$f_T(t)=\int_{-\infty}^{t}f_{Z_1,Z_2}(z_1,t-z_1)dz_1=\int_{-\infty}^{t}f_{Z_1,Z_2}(t-z_2,z_2)dz_2 \tag{5.19}$$

We already have the pdf and cdf of T from which we can compute the statistical value of T in order to test the hypothesis. The null hypothesis regarding the equality of the mean of two independent samples is therefore rejected if $T\leq t_0$ or $P\left(T\leq t_0\right)=F_T(t_0)\leq\alpha$.

The computation of $F_T(t_0)$ using Equation (5.17) involves the five summation and therefore rather difficult. The computation gets easier in the case where $\sigma_x^2 = \sigma_y^2$.

In the case of $\sigma_x^2 = \sigma_y^2$, we replace the V_x and V_y with $\frac{V_x+V_y}{2}$, therefore Equations (5.3) and (5.5) become

$$T^* = \frac{\frac{V_x+V_y}{2}}{\left[\frac{V_x+V_y}{2} + \frac{n(\overline{Y}-\overline{X})^2}{n-1}\right]} \tag{5.20a}$$

$$= \frac{U^*}{U^* + 4V^*} \tag{5.20b}$$

where $U^* = \frac{(n-1)\left(V_x+V_y\right)}{\sigma_x^2}$ and $V^* = \frac{n(\overline{Y}-\overline{X})^2}{2\sigma_x^2}$.

The pdf of T^* is derived from the ratio of linear combination of chi-square random variables [56]. First, let $Y = 1 + 4\frac{V^*}{U^*}$, where V^* is distributed $\chi^2_{(1)}$ and U^* is distributed $\chi^2_{2(n-1)}$. Second, the pdf of T^* is computed by taking $T^* = \frac{1}{Y}$.

In the following computation, the chi-square distribution is represented as the Gamma distribution. Therefore, we have V^* as Gamma distributed with parameters $\alpha_1 = \frac{1}{2}$ and $\beta_1 = 2$. U^* is Gamma distributed with parameters $\alpha_2 = (n-1)$ and $\beta_2 = 2$. U^* and V^* are independents.

Suppose $G = \frac{V^*}{U^*}$ and if we take $U^* = H$, then we get $V^* = GH$. Furthermore we have the Jacobian of this transformation random variable is h. Because V^* and U^* are independents, then the joint probability function of G and H is

$$f_{G,H}(g,h) = f_{V^*,U^*}(gh,h).h, \tag{5.21}$$

where

$$f_{V^*,U^*}(gh,h) = f_{V^*}(gh).f_{U^*}(h), \quad f_{V^*}(gh) = \frac{(gh)^{\alpha_1-1}e^{-\frac{gh}{\beta_1}}}{\beta_1^{\alpha_1}\Gamma(\alpha_1)}, \text{and} \quad f_{U^*}(h) = \frac{(h)^{\alpha_2-1}e^{-\frac{h}{\beta_2}}}{\beta_2^{\alpha_2}\Gamma(\alpha_2)}$$

Therefore

$$f_{G,H}(g,h) = \frac{(gh)^{\alpha_1-1}e^{-\frac{gh}{\beta_1}}}{\beta_1^{\alpha_1}\Gamma(\alpha_1)} \cdot \frac{(h)^{\alpha_2-1}e^{-\frac{h}{\beta_2}}}{\beta_2^{\alpha_2}\Gamma(\alpha_2)} h$$

$$= \frac{g^{\alpha_1-1}h^{\alpha_1+\alpha_2-1}e^{(-\frac{(1+g)}{\beta}h)}}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)}, \quad \beta_1 = \beta_2 = \beta. \tag{5.22}$$

To determine the pdf of g, then

$$\begin{aligned} f_G(g) =& \int_0^\infty f_{G,H}(g,h)dh \\ =& \frac{g^{\alpha_1-1}}{\mathrm{B}(\alpha_1,\alpha_2)(1+g)^{\alpha_1+\alpha_2}} \\ =& \frac{g^{\frac{1}{2}-1}}{\mathrm{B}(\frac{1}{2},n-1)(1+g)^{n-\frac{1}{2}}}, \end{aligned} \tag{5.23}$$

G is a beta of the second kind distribution.

The next step is determining the distribution of Y. We define $Y= 1+4G$ and by using the transformation random variable, where $G = \frac{Y-1}{4}$, the pdf of Y is computed as follows:

$$\begin{aligned} f_Y =& \frac{\left(\frac{y-1}{4}\right)^{\frac{1}{2}-1}}{\mathrm{B}(\frac{1}{2},n-1)\left(1+\left(\frac{y-1}{4}\right)\right)^{n-\frac{1}{2}}}\frac{1}{4} \\ =& \frac{4^{n-1}(y-1)^{\frac{1}{2}-1}}{\mathrm{B}(\frac{1}{2},n-1)(3+y)^{n-\frac{1}{2}}}, \quad 1 \le y \le \infty. \end{aligned} \tag{5.24}$$

The pdf of $T^* = \frac{1}{Y}$ is obtained from the following equations:

$$\begin{aligned} f_{T^*}(t^*) =& \frac{4^{n-1}(\frac{1}{t^*}-1)^{\frac{1}{2}-1}}{\mathrm{B}(\frac{1}{2},n-1)(3+\frac{1}{t^*})^{n-\frac{1}{2}}}\frac{1}{t^{*2}} \\ =& \frac{4^{n-1}t^{n-2}(1-t^*)^{\frac{1}{2}-1}}{\mathrm{B}(\frac{1}{2},n-1)(1+3t^*)^{n-\frac{1}{2}}}, \quad 0 \le t^* \le 1. \end{aligned} \tag{5.25}$$

Furthermore the cdf of T^* analytically is computed as

$$\begin{aligned} F_{T^*}(t_0^*) =& \int_0^{t_0^*} f_{T^*}(t^*)dt^* \\ =& \frac{4^{n-1}}{\mathrm{B}(\frac{1}{2},n-1)} \int_0^{t_0^*} \frac{t^{*(n-2)}(1-t)^{\frac{1}{2}-1}}{(1+3t^*)^{n-\frac{1}{2}}}dt^* \end{aligned}$$

$$
\begin{aligned}
=&\frac{4^{n-1}}{\mathrm{B}(\frac{1}{2},n-1)}\sum_{k=0}^{\infty}(-1)^k\binom{n-\frac{1}{2}+k-1}{k}\int_0^{t_0^*}t^{*(n-1+k-1)}(1-t^*)^{\frac{1}{2}-1}dt^*\\
=&\frac{4^{n-1}}{\mathrm{B}(\frac{1}{2},n-1)}\left[\sum_{k=0}^{\infty}(-1)^k\binom{n+k-\frac{3}{2}}{k}\mathrm{B}\left(t_0^*,n-1+k,\frac{1}{2}\right)\right],
\end{aligned}
\tag{5.26}
$$

where $B\left(t_0^*,n-1+k,\frac{1}{2}\right)=\int_0^{t_0^*}t^{*(n-1+k-1)}(1-t^*)^{\frac{1}{2}-1}dt^*$.

We will reject the null hypothesis of the equality of the mean of two independent samples if $T^*<t_{0,\alpha}^*$ or $P\left(t^*<T_0^*\right)=F_{T^*}(T_0^*)=\text{p-value}\leq P\left(t^*<t_{0,\alpha}^*\right)=F_{T^*}(t_{0,\alpha}^*)=\alpha$.

Observing that $(2(n-1))\left(\frac{V^*}{U^*}\right)$ is the square of a random variable having $t_{2(n-1)}$ distribution, a simple calculation shows that the same holds for the random variable

$$
J=\sqrt{(n-1)\left(\frac{1}{T^*}-1\right)} \tag{5.27}
$$

This statistic J can also be used to test the hypothesis $\mu_x=\mu_y$ and the critical values can be computed from the t table. It also follows that J has a limiting normal distribution as $n\to\infty$

5.3 The second proposed test

Consider that the distribution of T in Equation (5.5) is a sum of the dependent random variables Z_1 and Z_2. Because Z_1 and Z_2 are dependent then we need to compute the copula C_{Z_1,Z_2}. Calculating their functional copula are not easy, hence we approximate it by its empirical copula. For this, instead of using Equation (5.3), we use and rewrite Equation (2.3) as follows:

$$
\begin{aligned}
T=&\frac{1}{2}\left[\frac{V_x}{V_x+\frac{n(\overline{Y}-\overline{X})^2}{n-1}}+\frac{V_y}{V_y+\frac{n(\overline{Y}-\overline{X})^2}{n-1}}\right]\\
=&\left[\frac{V_x}{2V_x+2\frac{n(\overline{Y}-\overline{X})^2}{n-1}}\right]+\left[\frac{V_y}{2V_y+2\frac{n(\overline{Y}-\overline{X})^2}{n-1}}\right]\\
=&Z_1+Z_2.
\end{aligned}
\tag{5.28}
$$

Deheuvels (1979), in Genest and Favre [27] and Genest et al. [28], defines the empirical copula as follows:

Definition 5.2. Let $\{(Z_{1i}, Z_{2i}), i = 1, \ldots, n\}$ denote n independent observations of the vector (Z_1, Z_2). The empirical copula C_n is given by

$$C_n(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\widetilde{U}_1^i \leq u_1, \widetilde{U}_2^i \leq u_2), \tag{5.29}$$

where $\widetilde{U}_k^i = \frac{R_k^i}{n+1}$, $k = 1, 2$ are the components of the pseudo copula samples and $R_k^i = \sum_{j=1}^{n} \mathbf{1}(Z_k^j \leq Z_k^i)$ is the rank of the observation Z_k^i.

We simulate pseudo copula samples and observe that under the null hypothesis H_0 of equal means, the scatterplots show a linear dependence of the copula components and under the alternative hypothesis H_a the scatterplots show uncorrelated variables. See Figures 5.1, 5.2, and 5.3.

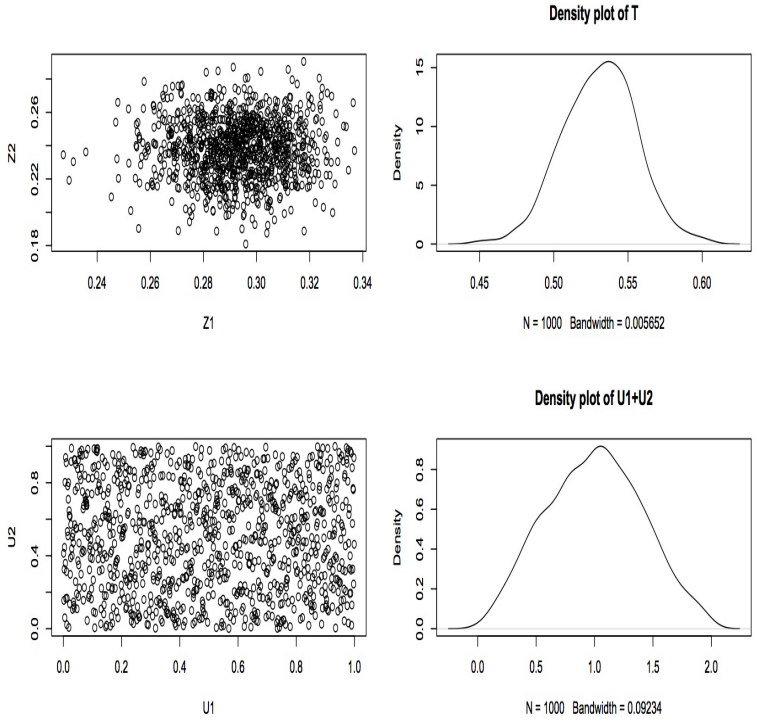


FIGURE 5.1: Supporting graphical plots in which the mean of two independent samples are different

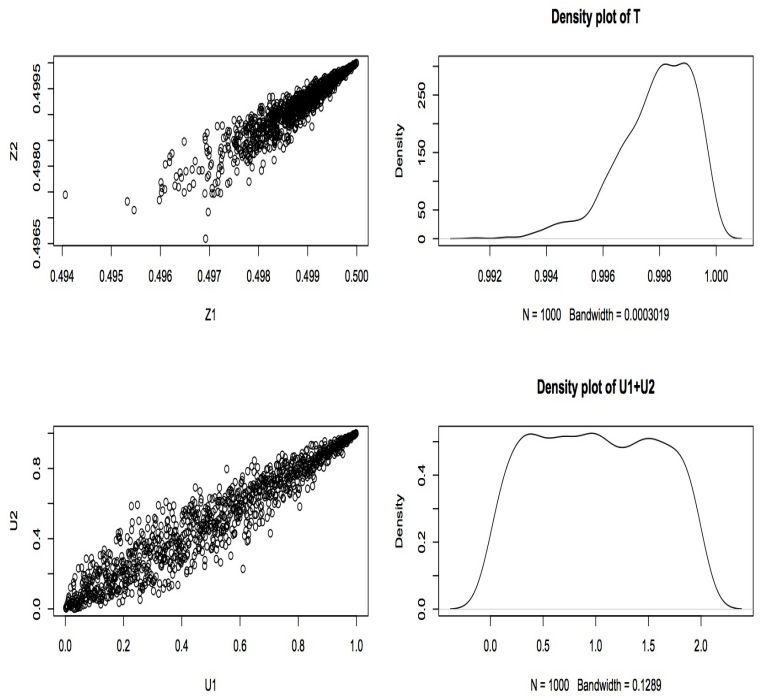


FIGURE 5.2: Supporting graphical plots in which the mean of two independent samples are equal

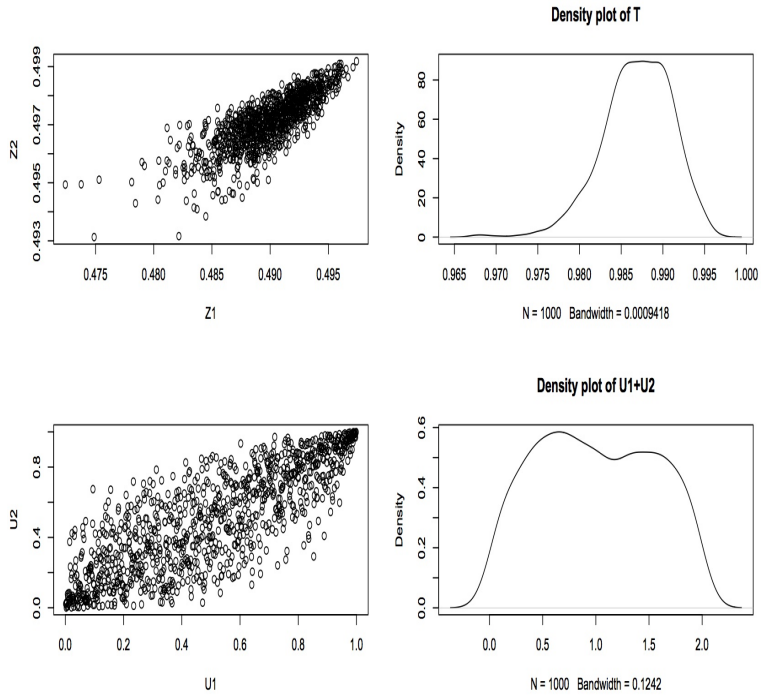


FIGURE 5.3: Supporting graphical plots, intermediate case

Thus it is natural to use the R^2 statistics to construct the test. We reject H_0 if $R^2 < \mathbf{c}$ where $\mathbf{c}$ is an α^{th} quantile of R_0^2 and R_0^2 is the R^2 under the null hypothesis.

Definition 5.3. Suppose we have n observations of two random variables, U_1 and U_2. The linear relationship between U_1 and U_2 is measured by the R^2, as follows:

$$R^2 = 1 - \frac{\frac{\sum_{i=1}^{n}(U_{2i}-\widehat{U}_{2i})^2}{n-2}}{\frac{\sum_{i=1}^{n}(U_{2i}-\overline{U}_{2})^2}{n-1}}, \tag{5.30}$$

where $\widehat{U}_{2i} = a + bU_{1i}$.

5.4 Simulation study

5.4.1 Simulation study the first proposed test, [21]

In this section we describe the results from the two simulation studies of the proposed test under the homogeneity of variance and the t test. First a simulation is conducted to measure the power of the proposed and the t tests, which is conducted at $N = 1000$ times and $\alpha = 0.01$. The results are presented in Section 5.4.1.1. In this simulation, we consider various possibilities regarding the sample sizes and variances. The results of the simulation are divided into groups according to

1. variance: low ($S = 0.2$), medium ($S = 1.2$) and high ($S = 7$),
2. sample size: low (5 and 25), medium (100) and high (500)

Section 5.4.1.2 describes the results of the second simulation measuring the rejection's rate under the null hypothesis of the proposed and the t tests, based on the $N = 500$ times simulation and $\alpha = 0.01$. In the simulation, we use $\mu_X = \mu_Y = \mu = 9.2$ and $\sigma_X = \sigma_Y = \sigma$ is chosen from these values of variances: $1.25, 3.5$ and 10 which represent the low, medium and high variance respectively.

Further details of both simulations are in Appendix F Section F.2.

5.4.1.1 Power of the test

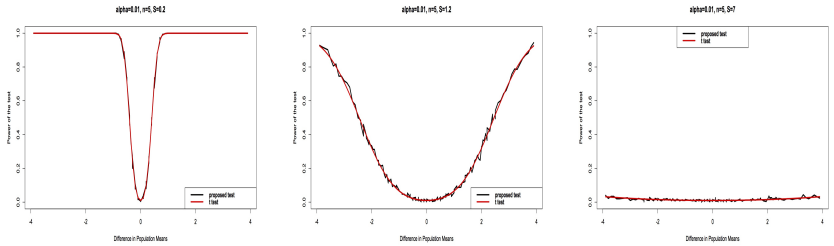


FIGURE 5.4: Graphical power of the t and proposed tests, n=5

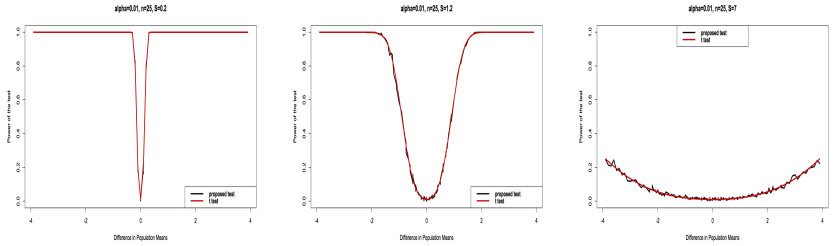


FIGURE 5.5: Graphical power of the t and proposed tests, n=25

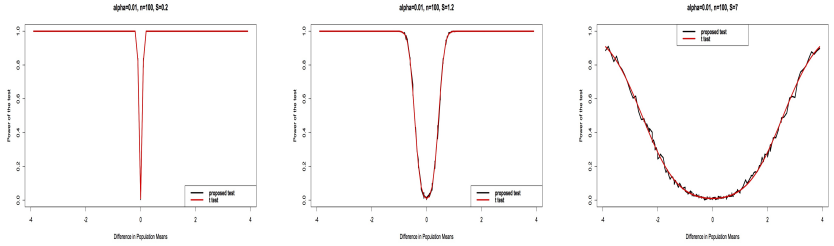


FIGURE 5.6: Graphical power of the t and proposed tests, n=100

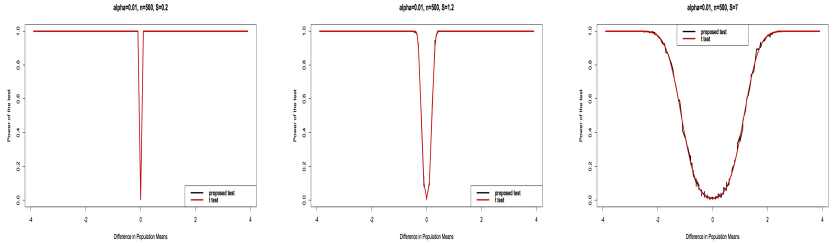


FIGURE 5.7: Graphical power of the t and proposed tests, n=500

Figures 5.4, 5.5, 5.6 and 5.7 present the power of the proposed and t tests, based on the simulation study.They show that the proposed and the t tests have an equal power. In the computation, the power of the proposed test is calculated using the empirical approach, therefore the result is not as smooth as the t test.

TABLE 5.1: Error type I rate under 500 simulation for the proposed and t tests

Sample size	Variance	proposed test		t test	
		0.05	0.01	0.05	0.01
5	low	0.056	0.012	0.056	0.012
	medium	0.062	0.016	0.062	0.016
	high	0.056	0.020	0.056	0.020
25	low	0.046	0.010	0.046	0.010
	medium	0.044	0.012	0.044	0.012
	high	0.038	0.010	0.038	0.010
100	low	0.058	0.006	0.058	0.006
	medium	0.050	0.010	0.050	0.010
	high	0.062	0.012	0.062	0.012
500	low	0.038	0.004	0.038	0.004
	medium	0.038	0.002	0.038	0.002
	high	0.050	0.010	0.050	0.010

5.4.1.2 Error type I

The simulation results of error type I are are shown in Table 5.1. It shows that the error type I rate of the t and the proposed tests are equal. This equality is also shown by the distribution of p-values in the proposed and t tests in Figures 5.8, 5.9 and 5.10.

(A) n=5 (B) n=500

FIGURE 5.8: P-values distribution of the proposed and t tests, small variance

(A) n=5 (B) n=500

FIGURE 5.9: P-values distribution of the proposed and t tests, medium variance

(A) n=5 (B) n=500

FIGURE 5.10: P-values distribution of the proposed and t tests, high variance

The next section provides some examples of comparison data analysis of the proposed and t tests. We used 14 artificial data sets of which the first 10 were taken randomly from the internet. The data can be seen in Table 5.2.

TABLE 5.2: Data sets, their mean and variance

Data	Data	Mean	Variance
1	X=(5,7,5,3,5,3,3,9)	5.000	4.571
	Y=(8,1,4,6,6,4,1,2)	4.000	6.571
2	X=(0.72,0.68,0.69,0.66,0.57,0.66,0.70,0.63,0.71,0.73)	0.675	0.002
	Y=(0.71,0.83,0.89,0.57,0.68,0.74,0.75,0.67,0.80,0.78)	0.742	0.008
3	X=(42,45,40,37,41,41,48,50,45,46)	43.500	15.833
	Y=(43,51,56,40,32,54,51,55,50,48)	48.000	57.330
4	X=(33,31,34,38,32,28)	32.667	11.067
	Y=(35,42,43,41)	40.250	12.917
5	X=(35,40,12,15,21,14,46,10,28,48,16,30,32,48,31, 22,12,39,19,25)	27.150	156.450
	Y=(2,27,38,31,1,19,1,34,3,1,2,3,2,1,2,1,3,29,37,2)	11.950	213.524
6	X=(26 21,22,26,19,22,26,25,24,21,23,23,18,29,22)	23.133	8.552
	Y=(18, 23, 21,20,20,29,20,16,20,26,21,25,17,18,19)	20.867	12.552
7	X=(520,460,500,470)	487.500	758.333
	Y=(230,270,250,280)	257.500	491.667
8	X=(3,0,6,7,4,3,2,1,4)	3.333	5.000
	Y=(5,1,5,7,10,9,7,11,8)	7.000	9.250
9	X=(16,20,21,22,23,22,27,25,27,28)	23.100	13.878
	Y=(19,22,24,24,25,25,26,26,28,32)	25.100	11.878
10	X=(91,87,99,77,88,91)	88.833	51.367
	Y=(101,110,103,93,99,104)	101.667	31.867
11	X=(10.11,7.36,6.34,11.83,8.61)	8.850	4.761
	Y=(3.28, 6.52,2.28,6.66,4.55)	4.658	3.760
12	X=(4.79,4.95,2.52,4.98,4.99)	4.446	1.166
	Y=(7.90,7.51,6.62,7.57,7.49)	7.418	0.227
13	X=(3.99,3.98,4.03,4.06, 3.84)	3.980	0.007
	Y=(6.68,6.25,6.97,5.75,4.01))	5.932	1.366
14	X=(10.16,8.26,16.23,1.44,0.66)	7.350	41.816
	Y=c(28.06,8.52,25.39,15.45,16.03)	18.690	63.422

TABLE 5.3: F.test decision

Samples	Decision	Samples	Decision
Data 1	equal variance	Data 8	equal variance
Data 2	equal variance	Data 9	equal variance
Data 3	equal variance	Data 10	equal variance
Data 4	equal variance	Data 11	equal variance
Data 5	equal variance	Data 12	equal variance
Data 6	equal variance	Data 13	not equal variance
Data 7	equal variance	Data 14	equal variance

TABLE 5.4: P-values and decisions from the proposed and t tests

Sample	t test		proposed test	
	p-value	Decision	p-value	Decision
Data 1	0.411	Accept	0.411	Accept
Data 2	0.054	Accept	0.054	Accept
Data 3	0.114	Accept	0.113	Accept
Data 4	0.009	Reject		
Data 5	0.001	Reject	0.001	Reject
Data 6	0.067	Accept	0.066	Accept
Data 7	0.000	Reject	0.000	Reject
Data 8	0.010	Accept	0.010	Accept
Data 9	0.229	Accept	0.229	Accept
Data 10	0.006	Reject	0.006	Reject
Data 11	0.012	Accept	0.012	Accept
Data 12	0.000	Reject	0.000	Reject
Data 14	0.039	Accept	0.039	Accept

TABLE 5.5: P-values and decision from the proposed test, Data set 4

Choice of n	Least Square	
	p-value	Decision
Min	0.021	Accept
Max	0.004	Reject
Average	0.009	Reject

5.4.1.3 Some examples

In this section, some example data sets are provided and used as an example of how to make the decision by using the special case of the cross variance test T^* (= the T test when $\sigma_X = \sigma_Y$). For this T^* test, the first step is making sure the variances of the samples are equal. There are some tests for this purpose, here we use the F.test in R.

In the proposed test, the null hypothesis is rejected if $t_o^* < t_\alpha^*$ or if the p-value of the special case of the cross variance test is less than α, where t_o^* is the t^* statistic

from the observed sample and p-value $= P\left(t_o^* < t_\alpha^*\right)$. In the computation we use $\alpha = 0.01$.

The data sets are shown in Table 5.2 and the computation results for the F.test are shown in Table 5.3. Tables 5.4 and 5.5 provide the results of the decision for both tests. Table 5.3 shows that data set 13 has unequal variances, therefore it will be excluded from further computation.

From Table 5.4 we can see that the p-values and decisions from the proposed and the t tests are equal, except for data set 4. It is the only data set with a different sample size. When the sample size of two samples is different, we provide two options of n: the $\max(n_1, n_2)$ or the average of (n_1, n_2). The example is shown in Table 5.5.

Based on the results of this section, we can suggest that the proposed test could be used as an alternative to detect whether or not there is difference between the means of two independent normal populations, where the variance between two populations is assumed to be equal.

5.4.2 Simulation study the second proposed test

In this section we describe the results from the simulation studies of the proposed and the t tests. The simulation is conducted to measure the power of both tests, with $M_1 = 1000$, $M_2 = 500$ and $\alpha = 0.05$.

In the simulation, we consider various possibilities regarding the sample sizes and variances. The results of the simulation are divided into groups according to

1. variance: low $(0.1 \leq S \leq 1.57)$, medium $(3.5 \leq S \leq 5.3)$ and high $(S \geq 6.97)$,
2. sample size (n): small $(5 \leq n \leq 10)$, medium $(15 \leq n \leq 25)$, large $(30 \leq n \leq 80)$ and very large $(n \geq 80)$

Further details of the simulation are provided in Appendix F Section F.3.

5.4.2.1 Power of the test

5.4.2.1.1 Homogeneous variance

When the variances are homogeneous and the variances are very small (the standard deviation = 0.1), then the proposed and the t tests have the same power, even when the sample size is as small as 5, see Figures 5.11 until 5.14. But when the standard deviation is more than 0.5, the power of the proposed test is better than the t test, unless the sample size is ≥ 50. This is shown at Figure 5.13.

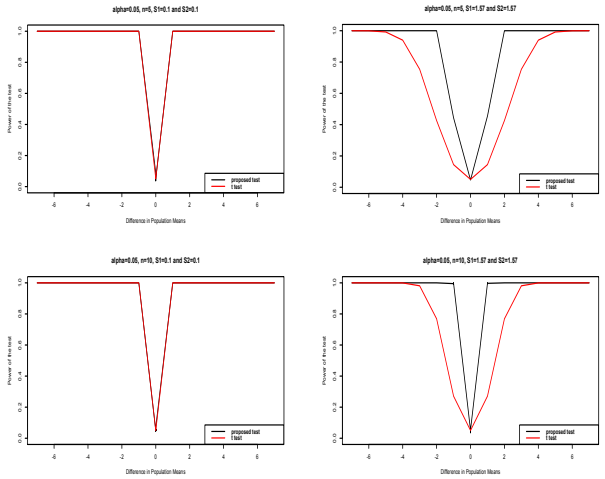


FIGURE 5.11: Graphical power of the proposed and the t tests, low homogeneity of variance (1)

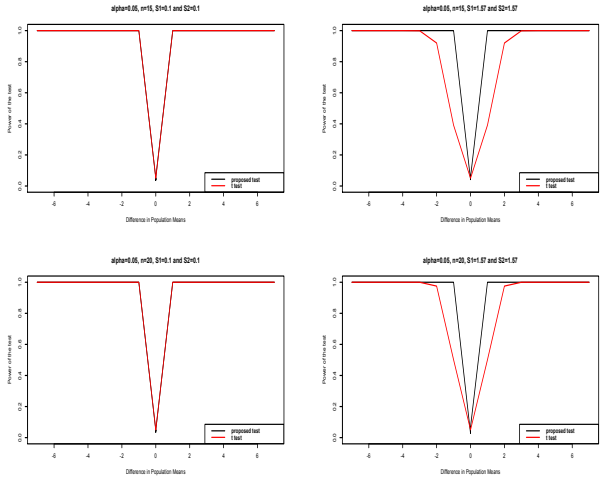


FIGURE 5.12: Graphical power of the proposed and the t tests, low homogeneity of variance (2)

When both of the variances are medium or high, in general the proposed test performs much better than the t test. Unless the sample size is really large, for example 500 as is shown in Figure 5.16. These are shown from Figures 5.15 to 5.18.

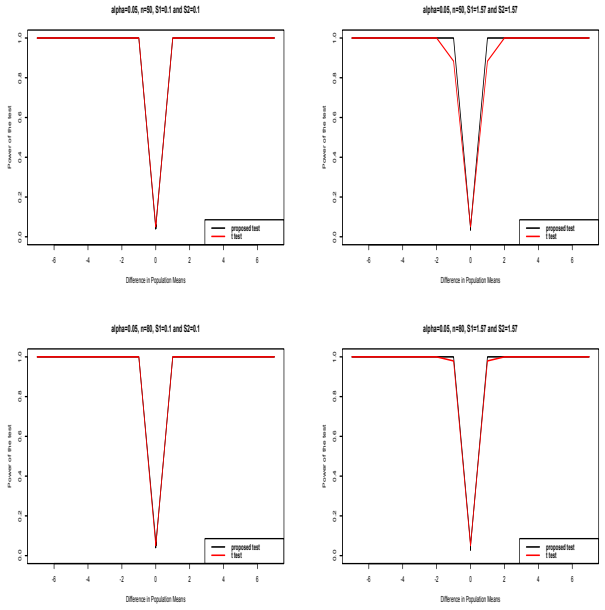


FIGURE 5.13: Graphical power of the proposed and the t tests, low homogeneity of variance (3)

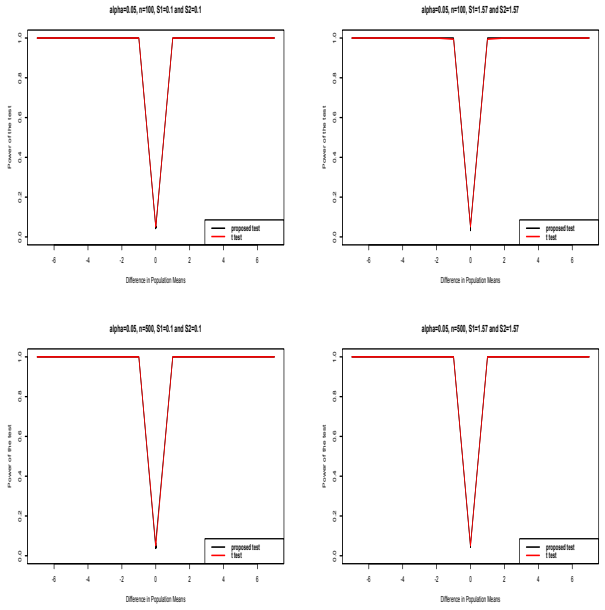


FIGURE 5.14: Graphical power of the proposed and the t tests, low homogeneity of variance (4)

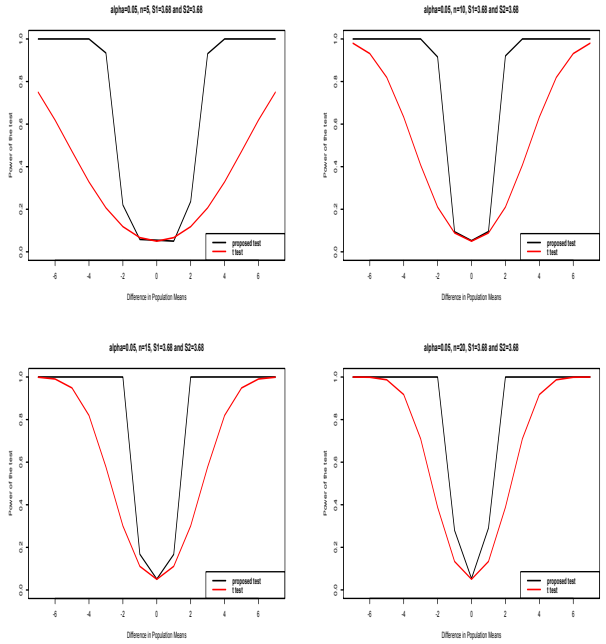


FIGURE 5.15: Graphical power of the proposed and the t tests, medium homogeneity of variance (1)

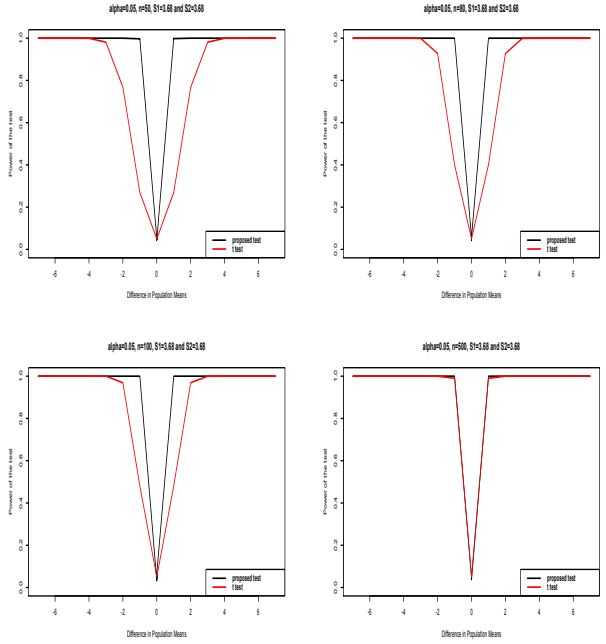


FIGURE 5.16: Graphical power of the proposed and the t tests, medium homogeneity of variance (2)

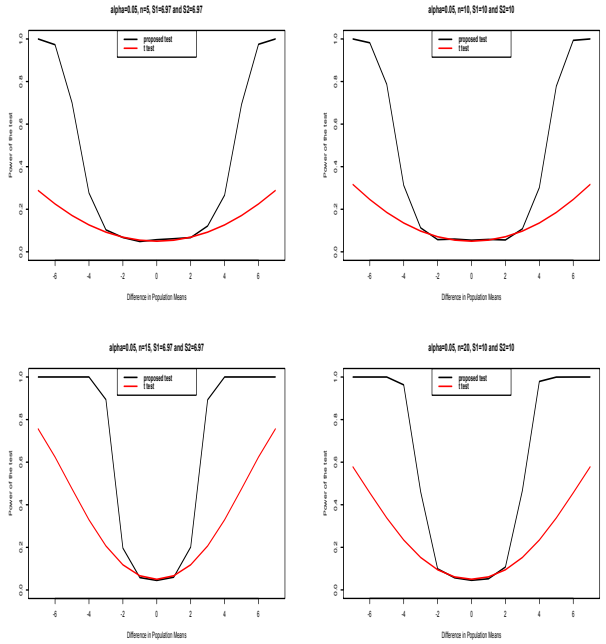


FIGURE 5.17: Graphical power of the proposed and the t tests, high homogeneity of variance (1)

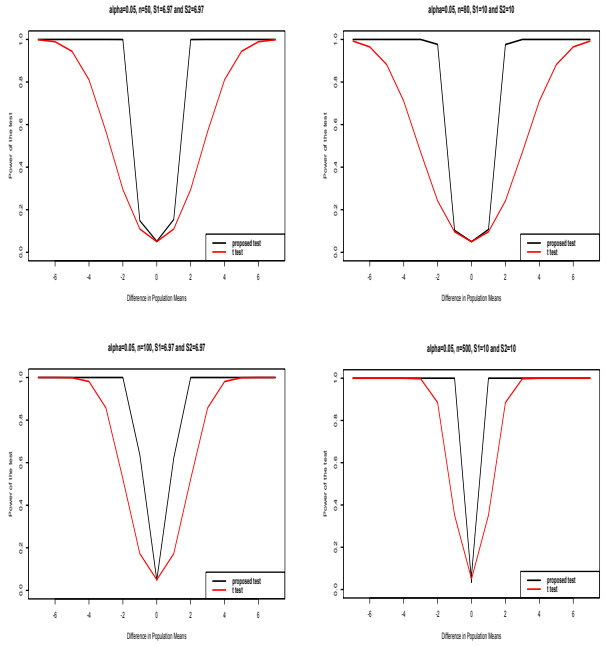


FIGURE 5.18: Graphical power of the proposed and the t tests, high homogeneity of variance (2)

5.4.2.1.2 Heterogeneous variance

When the variances are heterogeneous, the results are shown from Figures 5.19 to 5.25. Respectively, they describe the power when the variance of the two samples has a low, medium and high heterogeneity.

When the variance of the two samples is low and the heterogeneity between two variances is also low, then the power of the test could be equal if the sample size at least 15, otherwise the proposed test has higher power than the t test. See Figures 5.19 to 5.21.

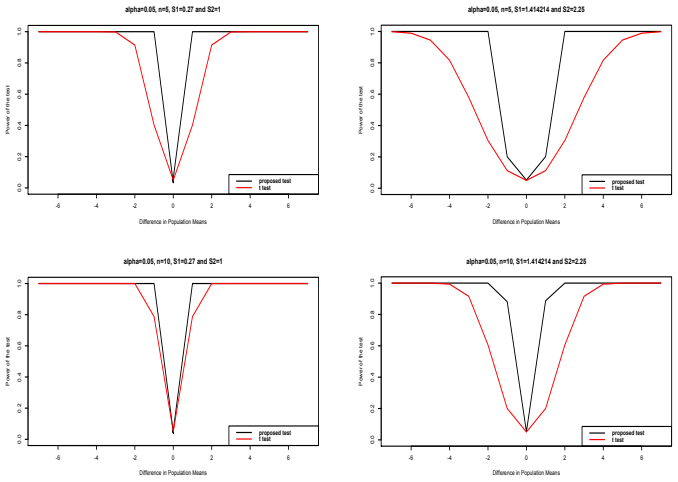


FIGURE 5.19: Graphical power of the proposed and the t tests, low heterogeneity of variance (1)

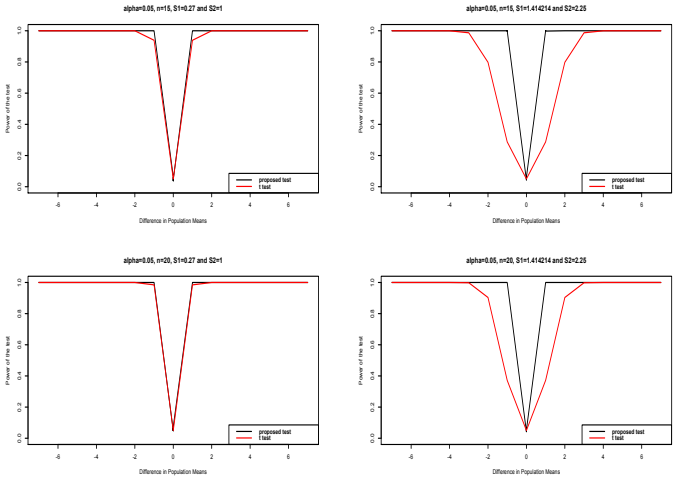


FIGURE 5.20: Graphical power of the proposed and the t tests, low heterogeneity of variance (2)

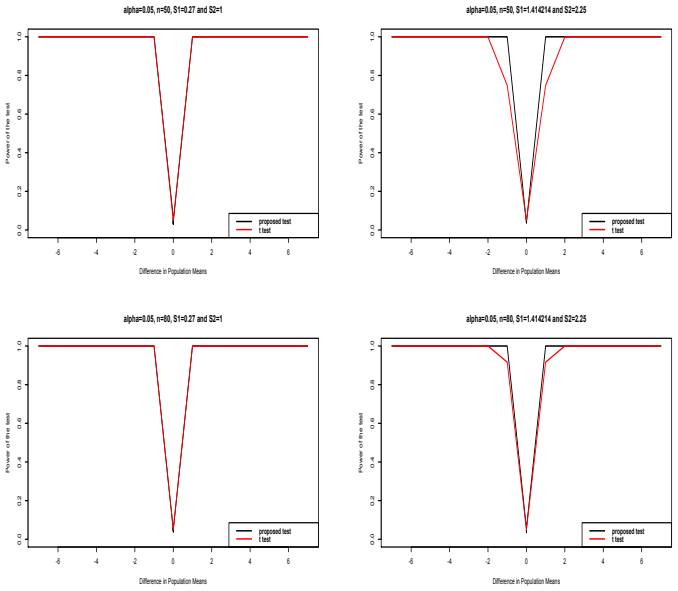


FIGURE 5.21: Graphical power of the proposed and the t tests, low heterogeneity of variance (3)

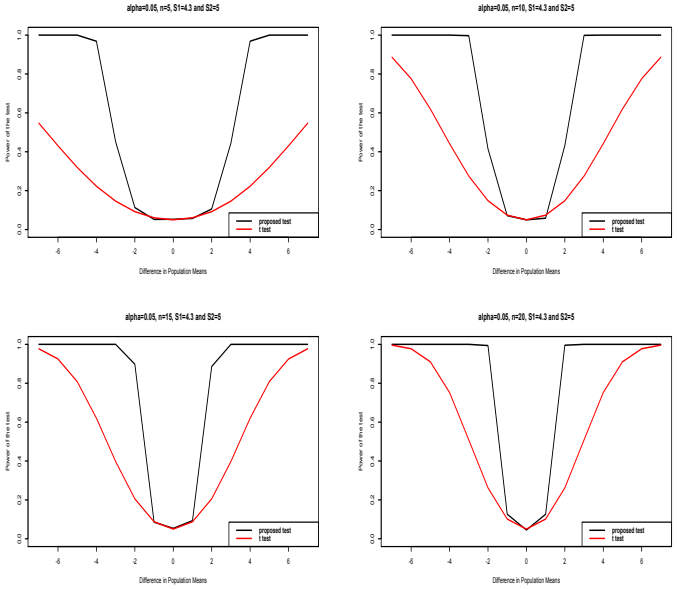


FIGURE 5.22: Graphical power of the proposed and the t tests, medium heterogeneity of variance (1)

In cases where the variances of the two samples are medium or high and the heterogeneity between two variances is also medium or high then Figures 5.22 to 5.27 show that the proposed test has higher power than the t test.

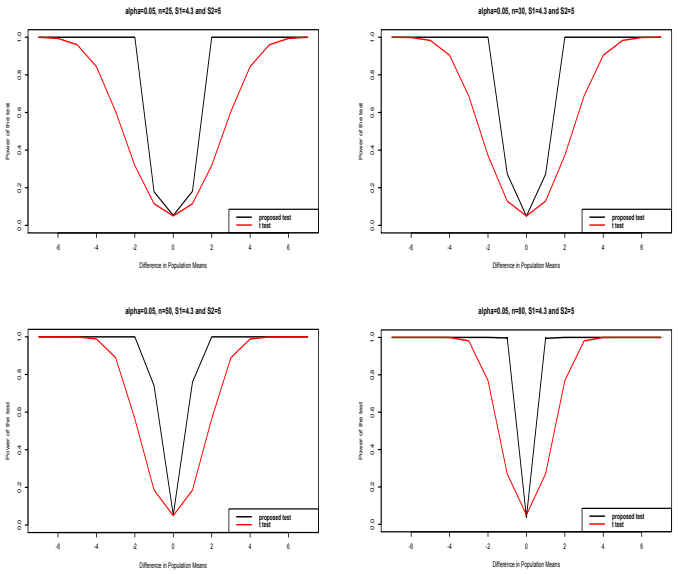


FIGURE 5.23: Graphical power of the proposed and the t tests, medium heterogeneity of variance (2)

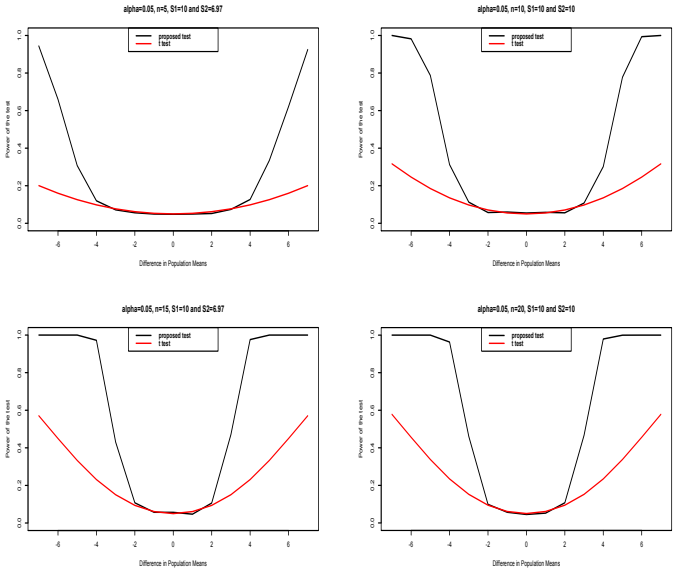


FIGURE 5.24: Graphical power of the proposed and the t tests, high heterogeneity of variance (1)

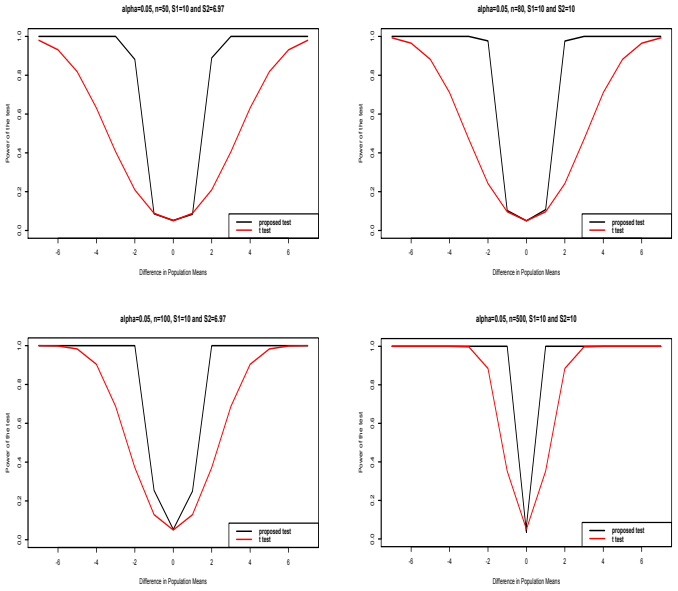


FIGURE 5.25: Graphical power of the proposed and the t tests, high heterogeneity of variance (2)

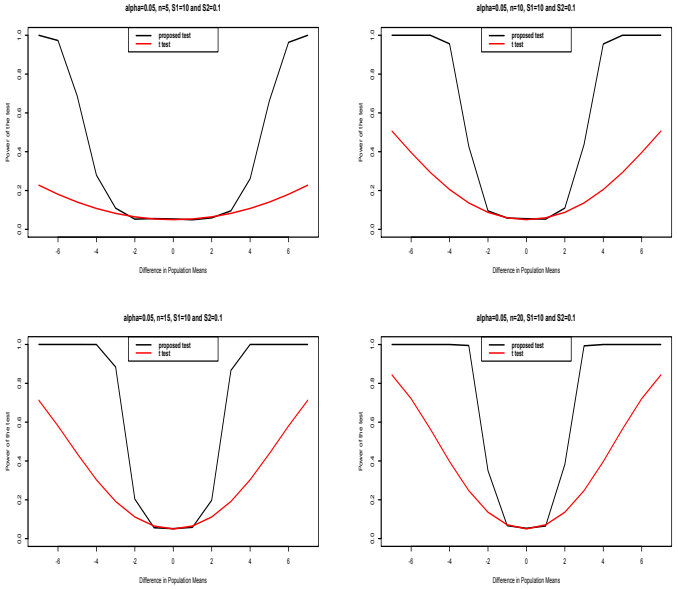


FIGURE 5.26: Graphical power of the proposed and the t tests, high heterogeneity of variance (3)

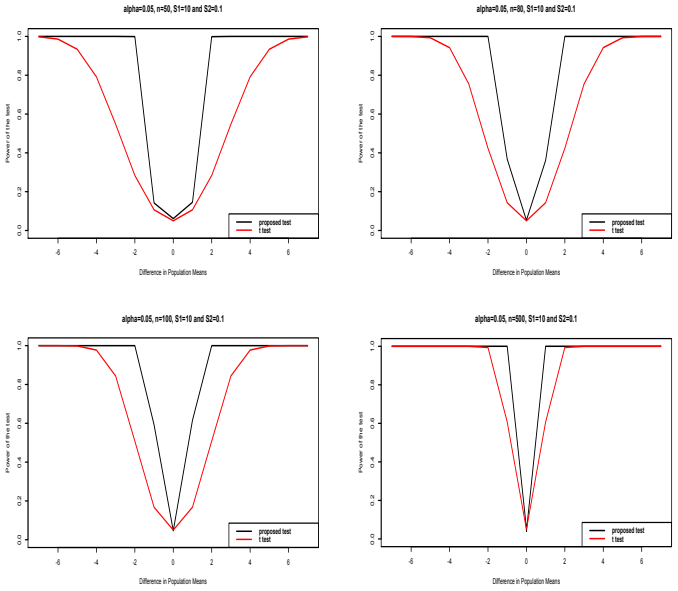


FIGURE 5.27: Graphical power of the proposed and the t tests, high heterogeneity of variance (4)

Some examples of data analysis concerning the proposed test and the comparison results with the t test will be described in the following section. The data sets are the same as in Table 5.2.

5.4.2.2 Some examples

In this section, some example data sets are provided and used to describe how to make the decision in the proposed test. The computation results are shown in Tables 5.6 and 5.7. We provided two options in rejecting or accepting the null hypothesis:

1. Use the graphs or plots

2. Use the A-value.
   Recall that the proposed test computation is based on generating M_2 of the empirical copula component samples and then compute the R^2 value. Before determining the R^2 distribution, we approximated the distribution under the null hypothesis by generating M_1 of the R^2 values.

   The proposed test provides M_1 of R_0^2 and R_1^2 values, for the null and the alternative hypotheses respectively. The A-value is computed as $\frac{\#(R_0^2 < \max(R_1^2))}{M_1}$.

In the computation we choose $M_1 = 1000$ for the graphics of the empirical copula components sample and $M_1 = 1000$ and $M_2 = 500$ for the A value.

TABLE 5.6: Decision based on graphical assessment

Samples	Decision	Explanation
Data 1	Accept H_0	The scatter plot or copula components are lying near a straight line and the plot of T shows left-skewed distribution
Data 2	Accept/Reject H_0	The scatter plot or copula components are lying near a straight line but some of them quite far from the ideal line. The plot of T shows almost right-skewed distribution although it is quite symmetric. One can accept or reject the null hypothesis.
Data 3	Accept/Reject H_0	equal to data set 2.
Data 4	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T shows right-skewed distribution
Data 5	Reject H_0	The scatter plot or copula components are lying randomly and the plot of T is right-skewed distribution are close to 0.4. In comparison, the data set 2 performs similary but the plot of T shows that the values are close to 1
Data 6	Accept/Reject H_0	equal to data set 2 and 3
Data 7	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T shows right-skewed distribution
Data 8	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line. The plot of T is right-slewed distribution.
Data 9	Accept H_0	The scatter plot or copula components are lying around a straight line and the plot of T shows left-skewed distribution
Data 10	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T is right-skewed distribution
Data 11	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T is right-skewed distribution
Data 12	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T is right-skewed distribution
Data 13	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T is right-skewed distribution
Data 14	Reject H_0	The scatter plot or copula components are lying randomly, far from a straight line and the plot of T is right-skewed distribution

5.4.2.2.1 The graphical assessment

All graphics for the data sets are in the Appendix Section G.2. The first option is implemented in 1000 simulations. Deciding that two means are equal means we are focusing on *Plot* T and *Plot* U_1 and U_2, which are at the top right and the bottom left in each graph respectively. The results are summarized in Table 5.6.

5.4.2.2.2 The A-value based assessment

In this computation, we need to reject the null hypothesis if A-value $\leq \alpha = 0.05$. All results from the computation are in Table 5.7.

From Table 5.7 we can see that the decisions from the proposed and the t tests are the same, except for the decision of data sets $2, 3$ and 6. However, if we used $\alpha = 0.01$ then the different decisions occur at the data sets $2, 3, 4, 6, 8, 11, 13$ and 14.

TABLE 5.7: p and A -values and decision from the t and the proposed tests

Sample	t test		proposed test	
	p-value	Decision	A-value	Decision
Data 1	0.412	Accept H_0	0.990	Accept H_0
Data 2	0.059	Accept H_0	0.000	Reject H_0
Data 3	0.119	Accept H_0	0.004	Reject H_0
Data 4	0.014	Reject H_0	0.000	Reject H_0
Data 5	0.001	Reject H_0	0.000	Reject H_0
Data 6	0.067	Accept H_0	0.000	Reject H_0
Data 7	0.000	Reject H_0	0.000	Reject H_0
Data 8	0.011	Reject H_0	0.000	Reject H_0
Data 9	0.229	Accept H_0	0.772	Accept H_0
Data 10	0.007	Reject H_0	0.000	Reject H_0
Data 11	0.013	Reject H_0	0.000	Reject H_0
Data 12	0.002	Reject H_0	0.000	Reject H_0
Data 13	0.020	Reject H_0	0.000	Reject H_0
Data 14	0.040	Reject H_0	0.000	Reject H_0

5.5 Remarks

The simulation study of the power of these tests shows that the first proposed and the t tests have the same power. Furthermore the p-value and the error type I rate under the null hypothesis of the proposed and t tests are exactly equal. These results suggest that the proposed model could be used as an alternative test for testing equality of mean of the two independent samples where the variance and sample sizes are equal.

The case studies of the example data sets show that the proposed test provides only one different decision from the t test, which happened when the sample sizes unequal. In this case, it is suggested to choose $\max(n_1, n_2)$ for the computation. We also introduce the new probability density functions which accompany the first proposed test.

The simulation study of the power of these tests shows that in some cases the second proposed and the t tests have equal power, although in general the proposed test has higher power than the t test.

The case studies of the example data sets show that the second proposed test provides some different decisions from the t test.

In view of the author, the second proposed method is suitable for data where replications of the samples are not many, such as in the bioinformatics field. However,

further investigation to compare the performance of this proposed test to other tests which are used in the bioinformatics, such as moderated t test [60], needs to be done to see which method leads to more experiment-wide power after false discovery rate control.

5.6 Chapter summary

In this chapter we introduced a cross-variance concept, two new tests based on the cross-variance to detect the difference in mean of two populations, and new probability density functions. A simulation study has been conducted to compute the power and the error type 1 rate of the proposed and the t tests.

The simulation study shows that the first proposed test performs equally to the t test. On the other hand the second proposed test performs better than the t test, proving that the second proposed test is more powerful than the t test. We believe that both of the the proposed tests could be used as an alternative to the t test.

Chapter 6

Conclusions and indication of future work

6.1 Conclusions

We have studied the additive models of background correction for the Illumina BeadArrays and proposed some new models where the true intensity is assumed to have exponential or gamma distribution and the noise is lognormally distributed. We have derived the formula of the true intensity value of the proposed models.

Furthermore, we compared the performance of all models, based on the benchmarking and public data sets. It has been shown that the proposed models perform moderately better than the existing models. We also proposed a generalized model where the true intensity is skewed-distributed and the noise can be skewed or symmetrical. In all proposed models we derived the formula for the true intensity value.

We introduced and developed alternative tests to detect difference of mean between two independent samples. The tests are based on the cross variance concept, pseudo copula samples, R^2 and bootstrapping. Although our tests offer some advantages, there are things that should be addressed in future work.

6.2 Future work

In the future we will address the following issues:

1. the generalized model of the background correction of the Illumina BeadArrays has not yet been fully implemented and applied
2. the implementation of the empirical copula as a dependence measure has not yet been explored
3. the theoretical part of the second proposed test needs to be investigated further

Appendix A

MA plots

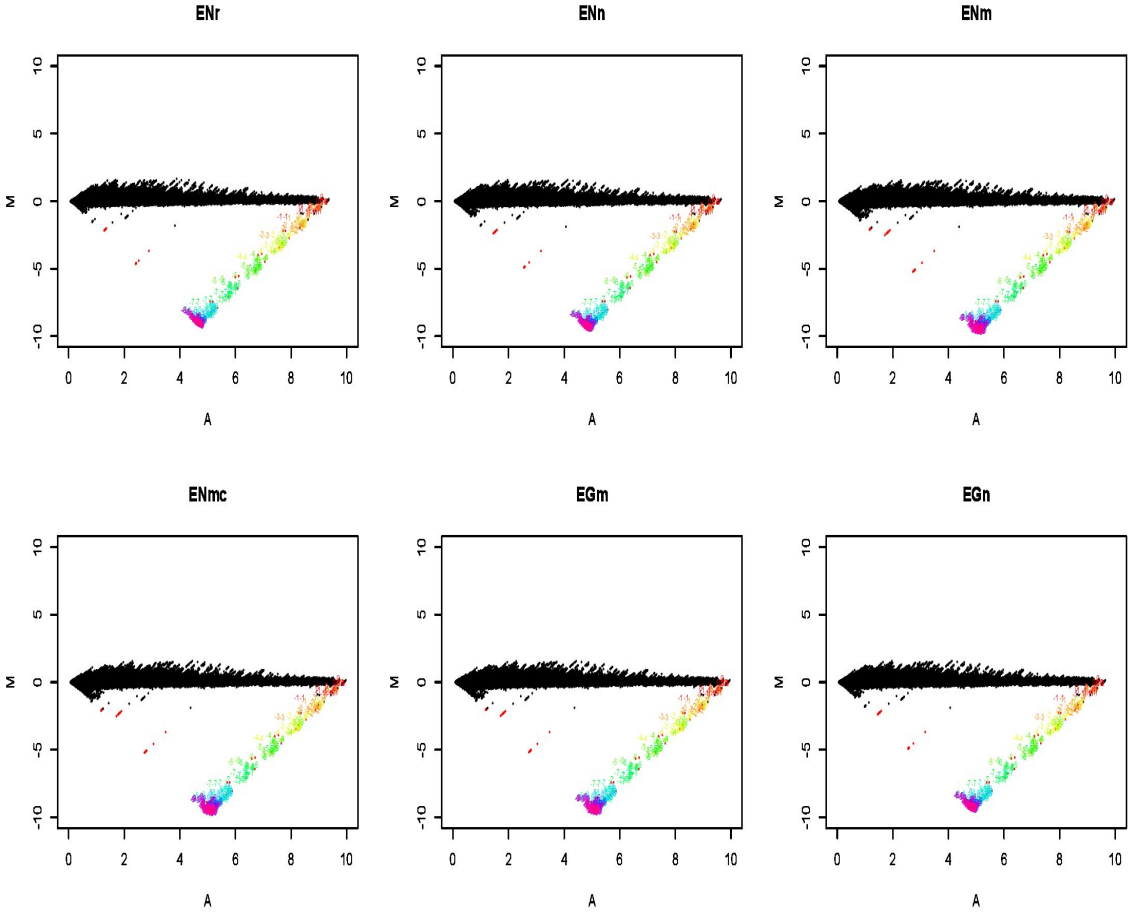


FIGURE A.1: MA plots. *(cont.)*

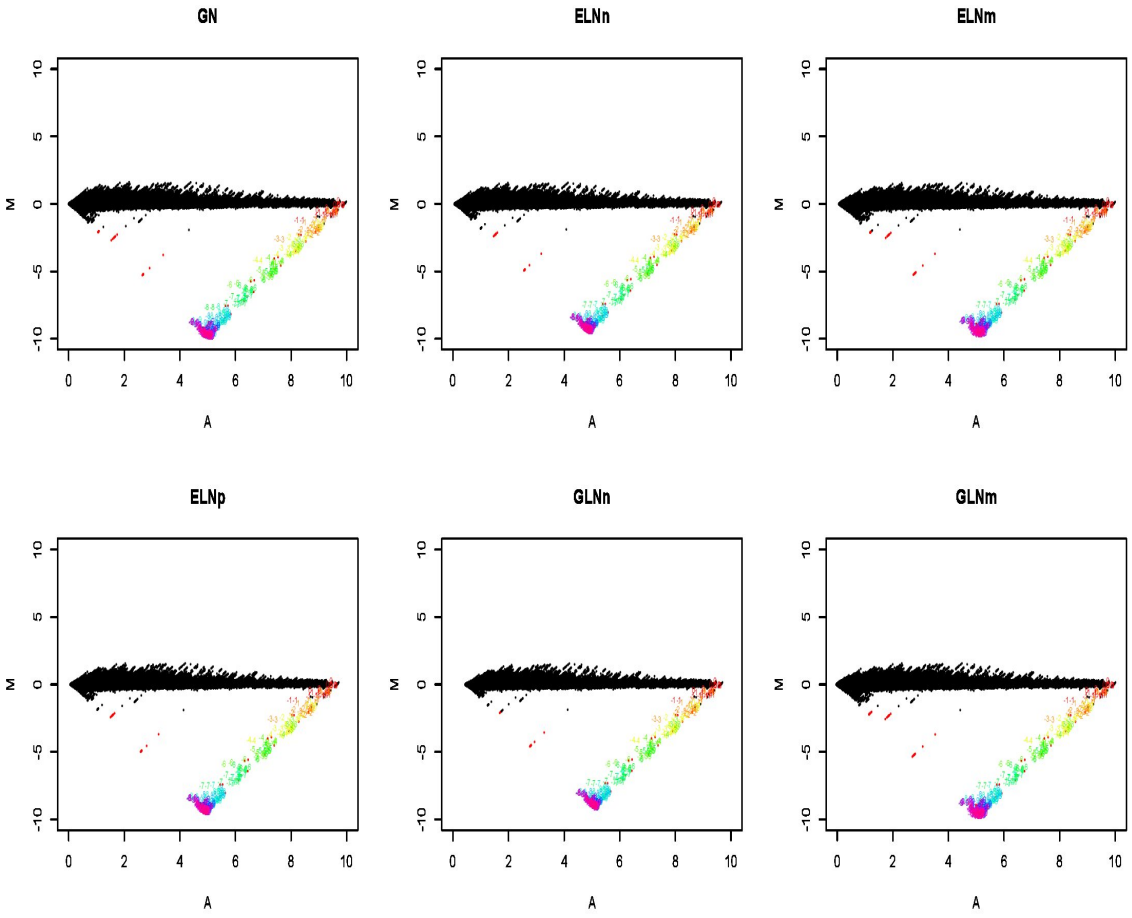


FIGURE A.2: MA plots. *(cont.)*

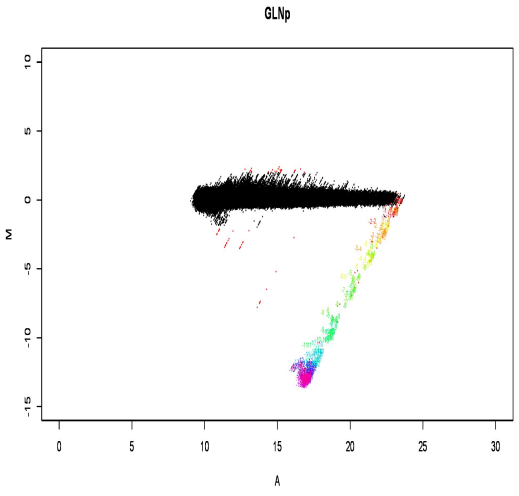


FIGURE A.3: MA plots.

Appendix B

Variance across replicates

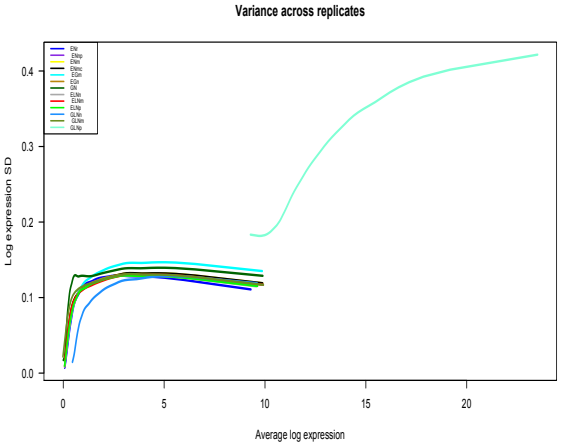


FIGURE B.1: Variance across replicates plots, all models

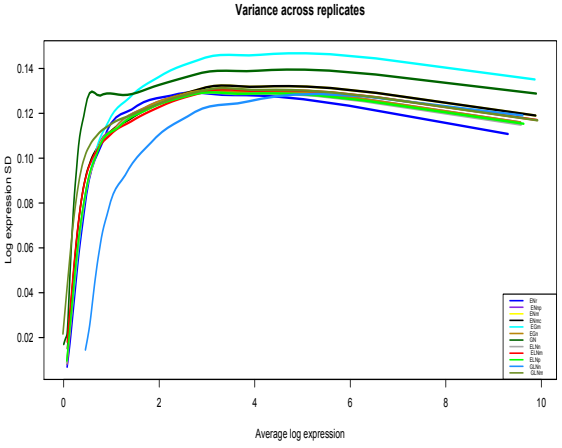


FIGURE B.2: Variance across replicates plots, without GLNp.

Appendix C

Nominal vs observed intensity

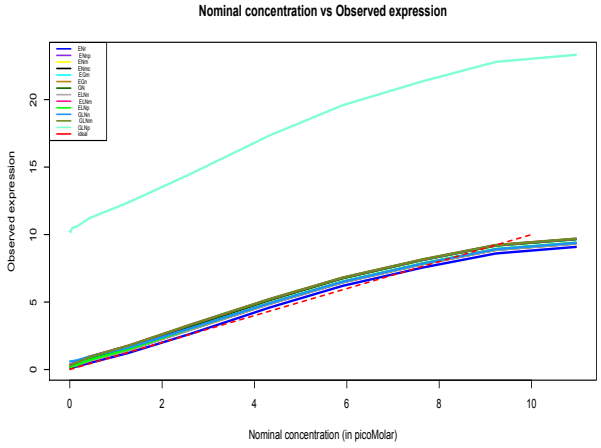


FIGURE C.1: Nominal vs observed plots, all models

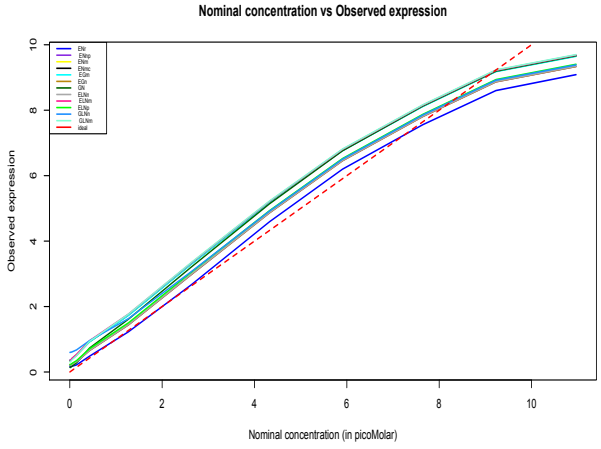


FIGURE C.2: Nominal vs observed plots, without GLNp.

Appendix D

Nominal vs observed fold-change

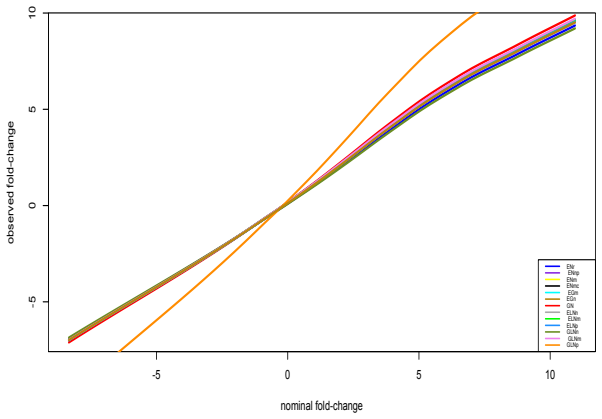


FIGURE D.1: Nominal vs observed fold-change plots, all arrays

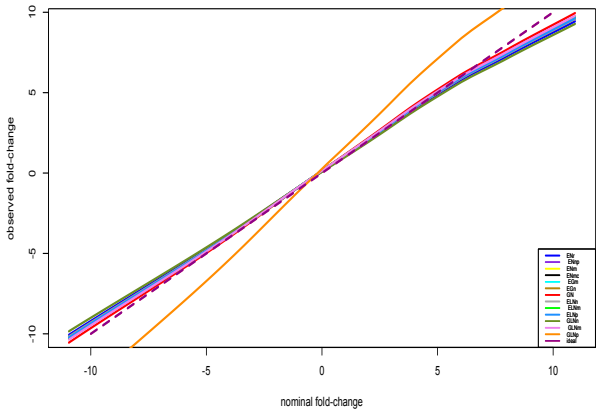


FIGURE D.2: Nominal vs observed fold-change plots, 24 arrays.

Appendix E

ROC curves

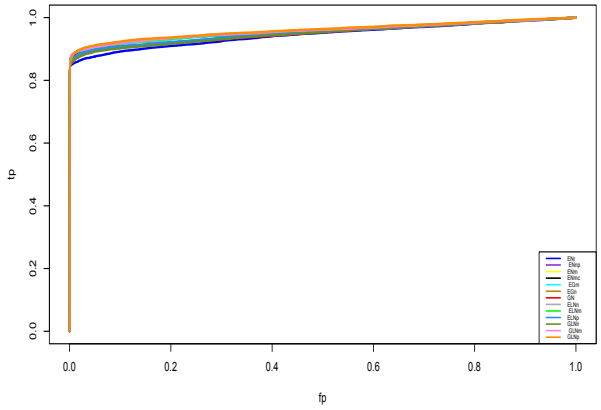


FIGURE E.1: ROC plots, all models

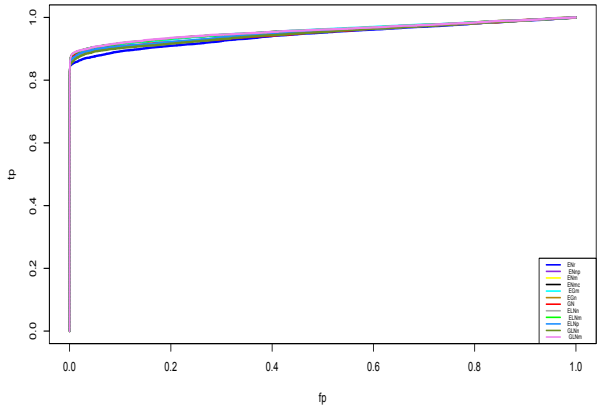


FIGURE E.2: ROC plots without GLNp

Appendix F

Simulation study

F.1 Benchmarking study

Let N be the number of simulations, n_1 the sample size of regular probes, n_2 the sample size of negative probes, Θ the original parameter vector of the underlying distribution in each model from the data set and $\hat{\Theta}$ the parameter vector of the underlying distribution in each model from the simulation data. In this dissertation we chose $N = 100$, $n_1 = 25000$ and $n_2 = 1000$.

The simulation is conducted by referring to the convolution in Equation (2.1), based on the underlying distribution. Once we have chosen the model, then

1. choose the parameters for the simulation. The parameters for the simulation are a combination of the minimum, median and maximum values of the original parameters. The original parameters are estimated from the data set based on the chosen model.

2. generate a sample for the true intensity (S), negative probes (B) and regular probes ($S + B$)

3. estimate the parameters of the underlying distribution based on the generated sample (Θ) and save

4. compute the true intensity value ($\hat{S}$) and save

5. repeat the steps above N times, then

6. compute the simulation criteria, to measure the bias of the background correction and the parameters:

   (a) MSE_{bc} is defined as

   $$\text{MSE}_{bc} = \frac{1}{N} \sum_{l=1}^{N} \left(\frac{1}{n_1} \sum_{j=1}^{n_1} \left(\hat{S}(P_j^l | \hat{\Theta}_l) - S_j^l \right)^2 \right)$$

   (b) L_1 error is defined as

   $$L_1 = \frac{1}{N} \sum_{l=1}^{N} \frac{\mid \Theta - \hat{\Theta}_l \mid}{\Theta}$$

The lower these MSE_{bc} and L_1 errors are, the better the signal can be estimated under such a model.

F.2 First proposed test

F.2.1 Simulation to compute the rejection rate under the null hypothesis between the proposed and the t tests

The simulation study is conducted as follows:

1. Choose the $n, M_1, \alpha, \mu_X = \mu_Y = \mu$ and $\sigma_X = \sigma_Y = \sigma$ of the two-groups independent samples
2. Repeat M_1 times
   (a) Compute p-value of t_o^* test
   (b) Compute p-value of t test
3. Compute the proportion of t test p-value ¡ α in M_1 results
4. Compute the proportion of the proposed test p-value ¡ α in M_1 results
5. Compare the result of steps 3 and 4

F.2.2 Simulation to compute the power of the t and the proposed tests 2

The simulation study is conducted to show the power of the second proposed test. It is implemented as follows:

1. Choose the $\mu_X, \mu_Y, \sigma_X = \sigma_Y = \sigma$ of the two-groups independent samples
2. The simulation under null hypothesis
   (a) Generate n random samples normally distributed with mean and standard deviation, μ_{X_0} and σ
   (b) Generate n random samples normally distributed with mean and standard deviation, μ_{Y_0} and σ
   (c) Compute their mean and variance samples
   (d) Compute t_0^* values, based on the equation (5.20a).

3. The simulation under alternative hypothesis

   (a) Generate n random samples normally distributed with mean and standard deviation, μ_{X_1} and σ

   (b) Generate n random samples normally distributed with mean and standard deviation, μ_{Y_1} and σ

   (c) Compute their mean and variance samples

   (d) Compute t_1^* values, based on the equation (5.20a).

4. Repeat steps (2) - (3), M times

5. Compute $t_{0,\alpha}^*$, the α quantile of t_0^*. Note that $t_{0,\alpha}^*$ can also be computed by using the α^{th} quantile of pdf of T^* in the Equation (5.25).

6. Compute the power of the proposed test $= 1 - \frac{\text{sum}(t_1^* \geq t_{0,\alpha}^*)}{M}$

7. Compute the power of t test from samples X_1 and Y_1

8. Compare the results from steps (6) and (7)

9. Do steps (1)-(8), for different values of mean and variance

F.3 Second proposed test

Simulation to compute the power of the proposed and the t tests 2

The simulation study is conducted to show the power of the proposed and the t tests. It is implemented as follows:

1. Choose the μ_X, μ_Y, σ_X and σ_Y of the two-groups independent samples

2. The simulation under null hypothesis

   (a) Generate n random samples normally distributed with mean and standard deviation, μ_{X_0} and σ_{X_0}

   (b) Generate n random samples normally distributed with mean and standard deviation, μ_{Y_0} and σ_{Y_0}

   (c) Compute their mean and variance samples

   (d) Compute Z_{01} and Z_{02} values, based on the equation (5.27).

3. The simulation under alternative hypothesis

   (a) Generate n random samples normally distributed with mean and standard deviation, μ_{X_1} and σ_{X_1}

   (b) Generate n random samples normally distributed with mean and standard deviation, μ_{Y_1} and σ_{Y_1}

   (c) Compute their mean and variance samples

   (d) Compute Z_{11} and Z_{12} values, based on the equation (5.27).

4. Repeat steps (2) - (3), M_2 times

5. Rank the values of Z_{01}, Z_{02}, Z_{11} and Z_{12} and divide by $M_1 + 1$

6. Compute the R_0^2 and R_1^2

7. Repeat steps (2) - (5), M_1 times

8. Compute R_0, the α quantile of R_0^2

9. Compute the power of the proposed test $= 1 - \frac{\text{sum}(R_1^2 \geq R_0)}{M_1}$

10. Compute the power of t test from samples X_1 and Y_1

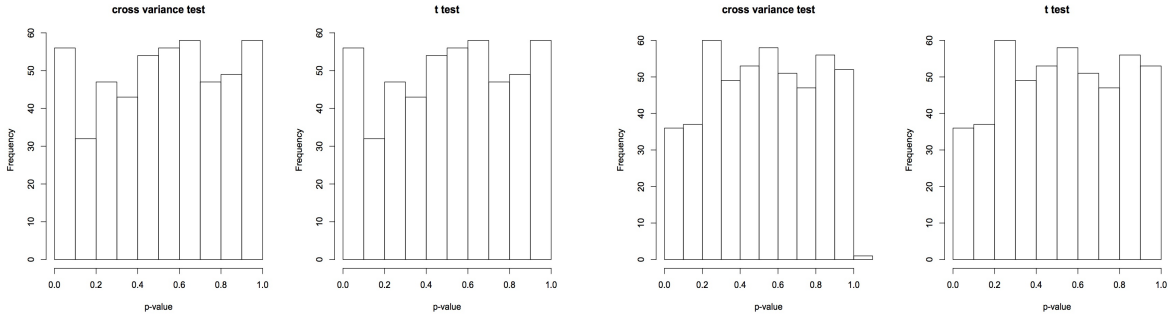
11. Compare the results from steps (9) and (10)

12. Do steps (1)-(11), for different mean and variance

Appendix G

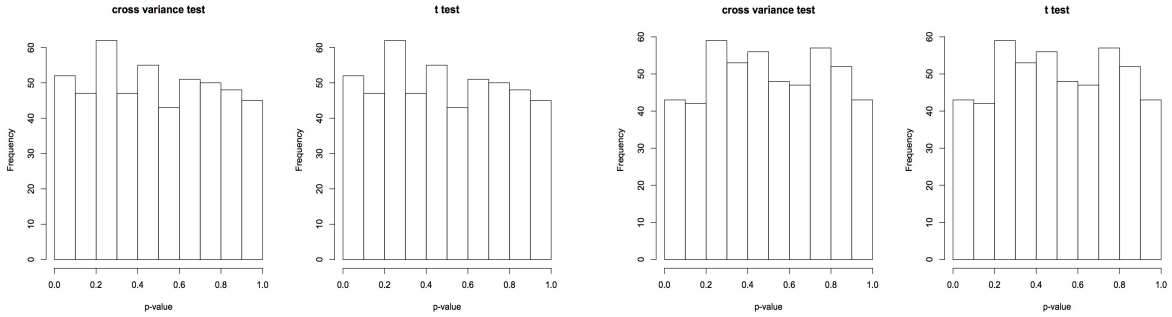
Supplementary graphs

G.1 First proposed test: P-values distribution



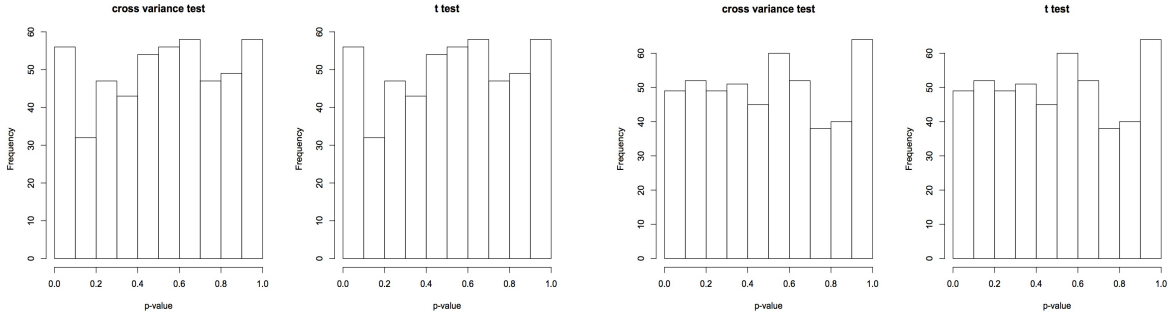
(A) n=5 (B) n=500

FIGURE G.1: P-values distribution of the proposed and t tests, small variance



(A) n=5 (B) n=500

FIGURE G.2: P-values distribution of the proposed and t tests, medium variance



(A) n=5 (B) n=500

FIGURE G.3: P-values distribution of the proposed and t tests, high variance

G.2 Second proposed test: Graphics of example data

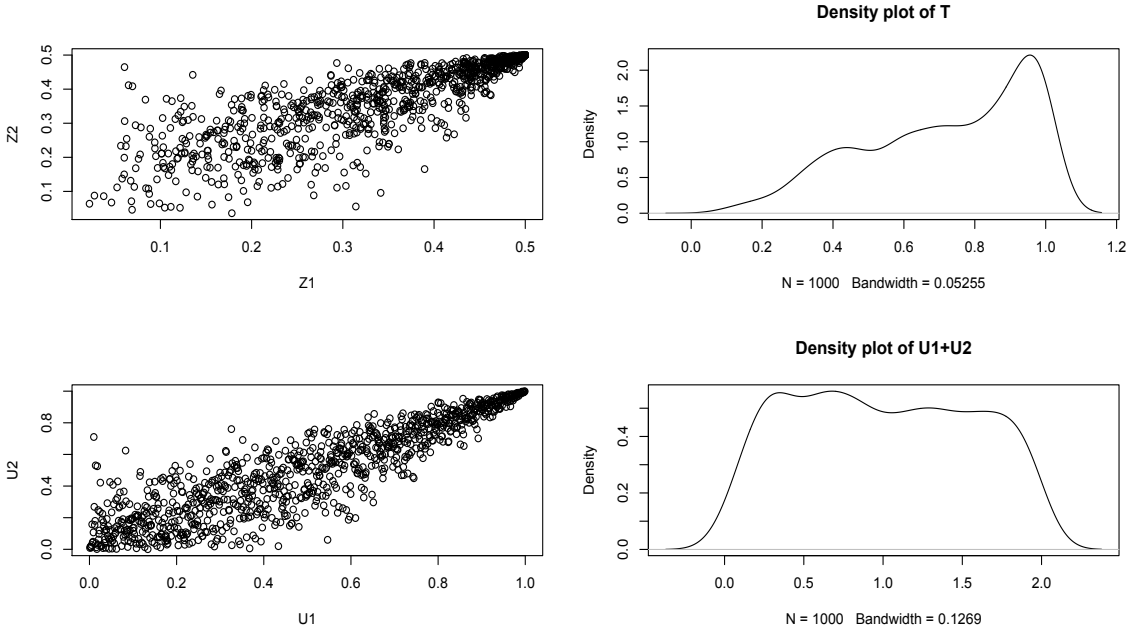


FIGURE G.4: Graphics of data set 1

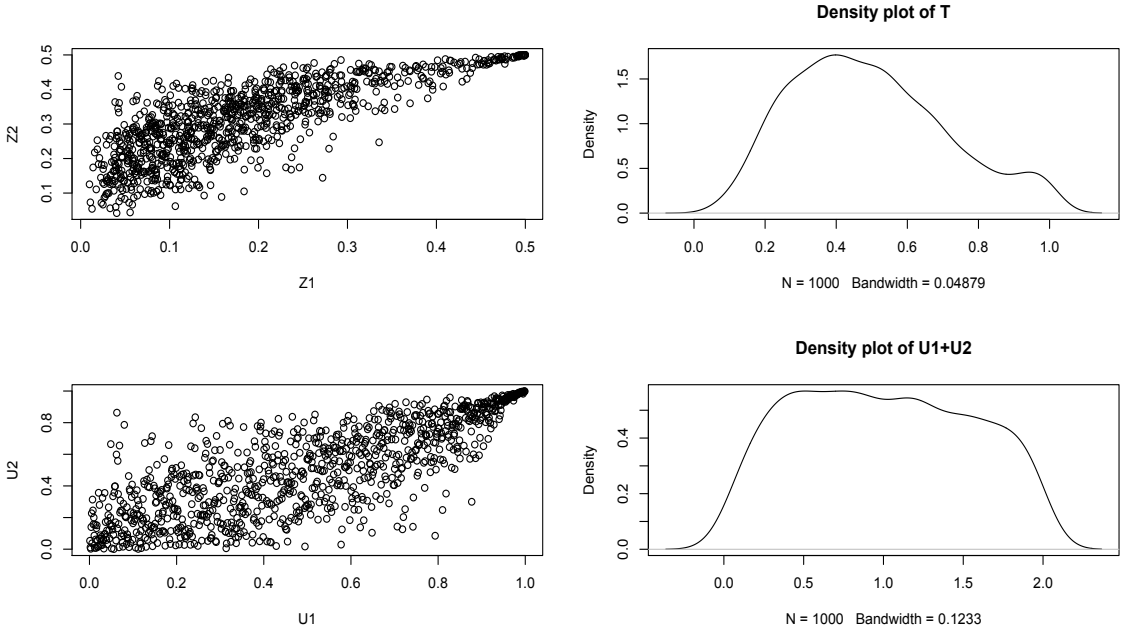


FIGURE G.5: Graphics of data set 2

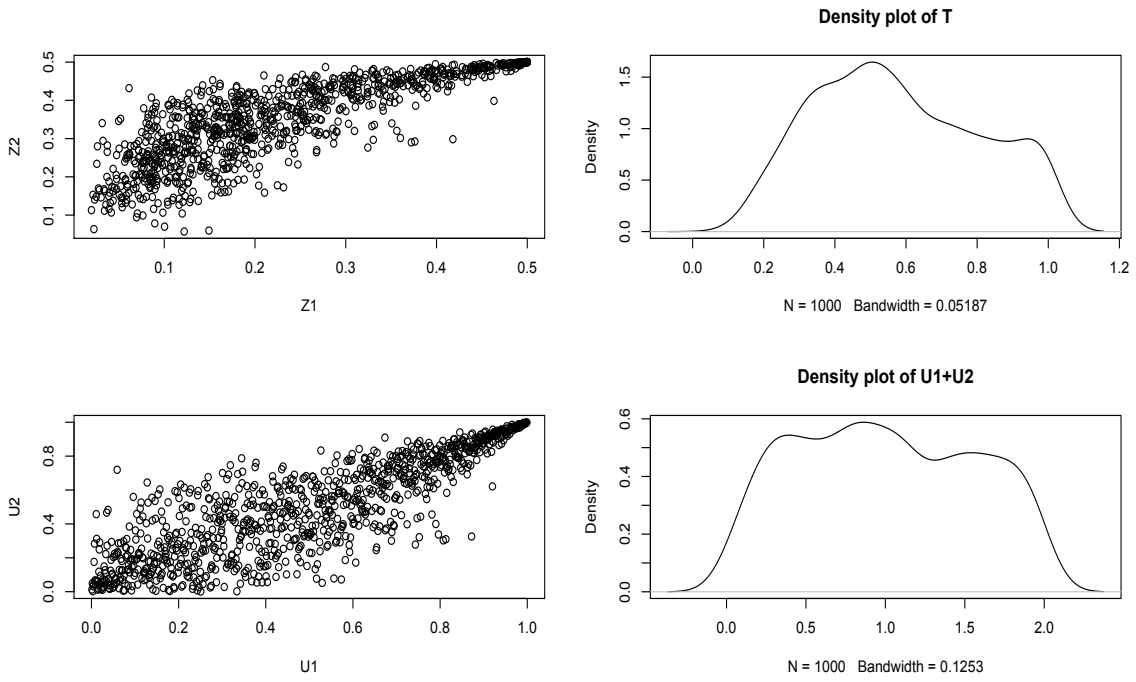


FIGURE G.6: Graphics of data set 3

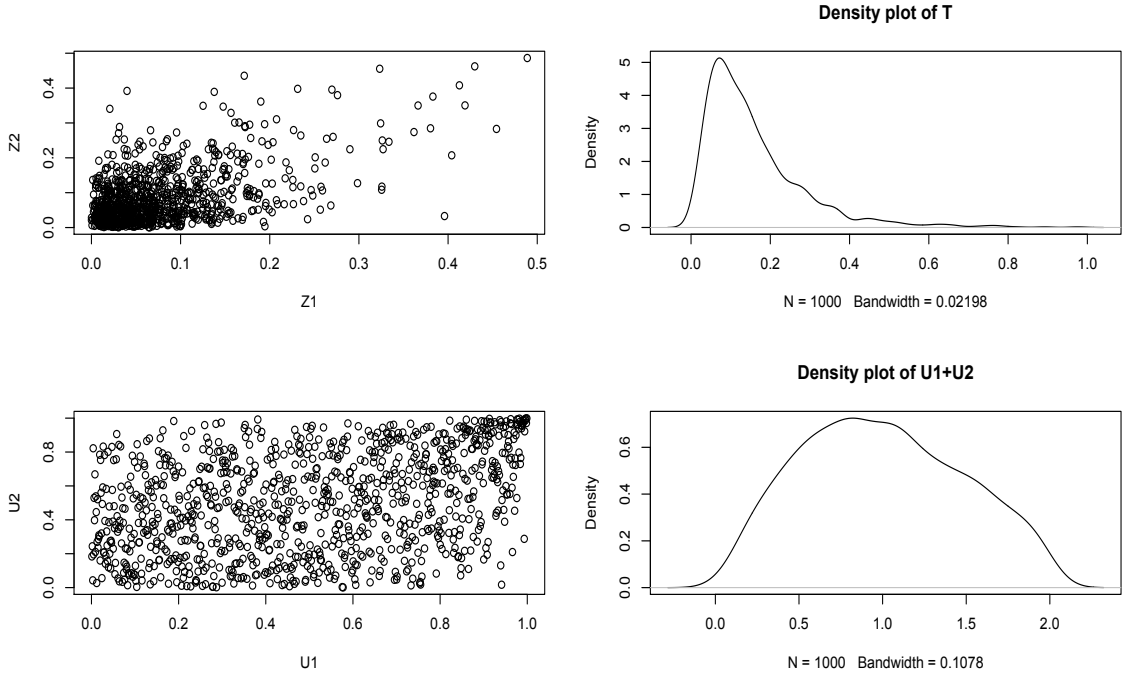


FIGURE G.7: Graphics of data set 4

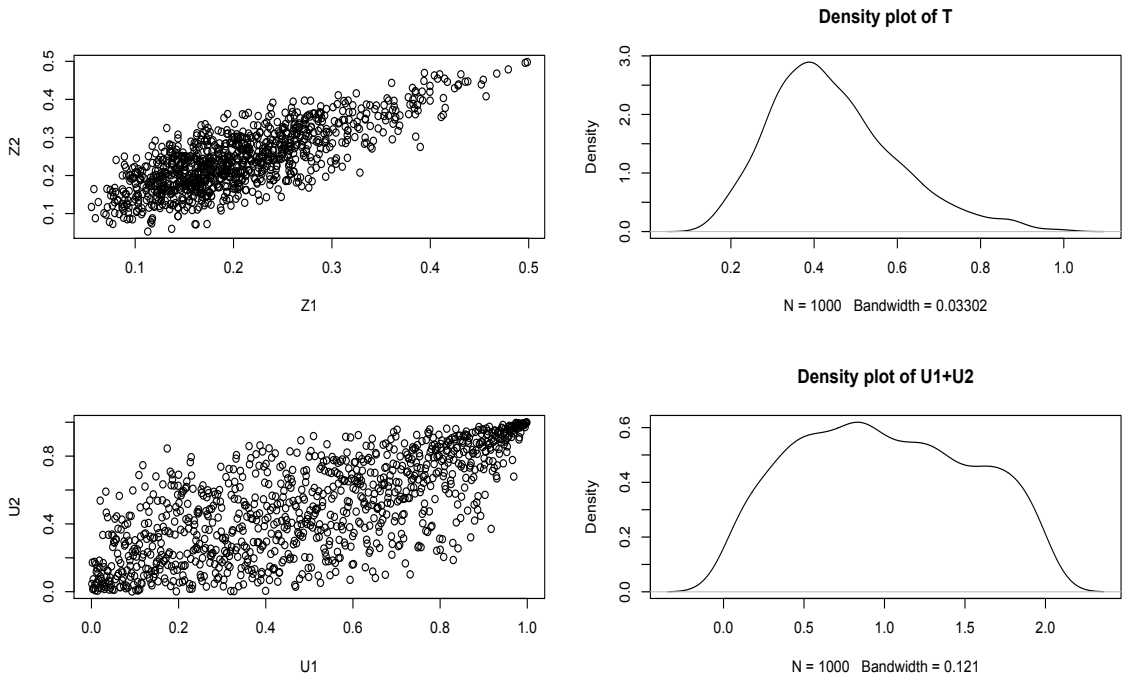


FIGURE G.8: Graphics of data set 5

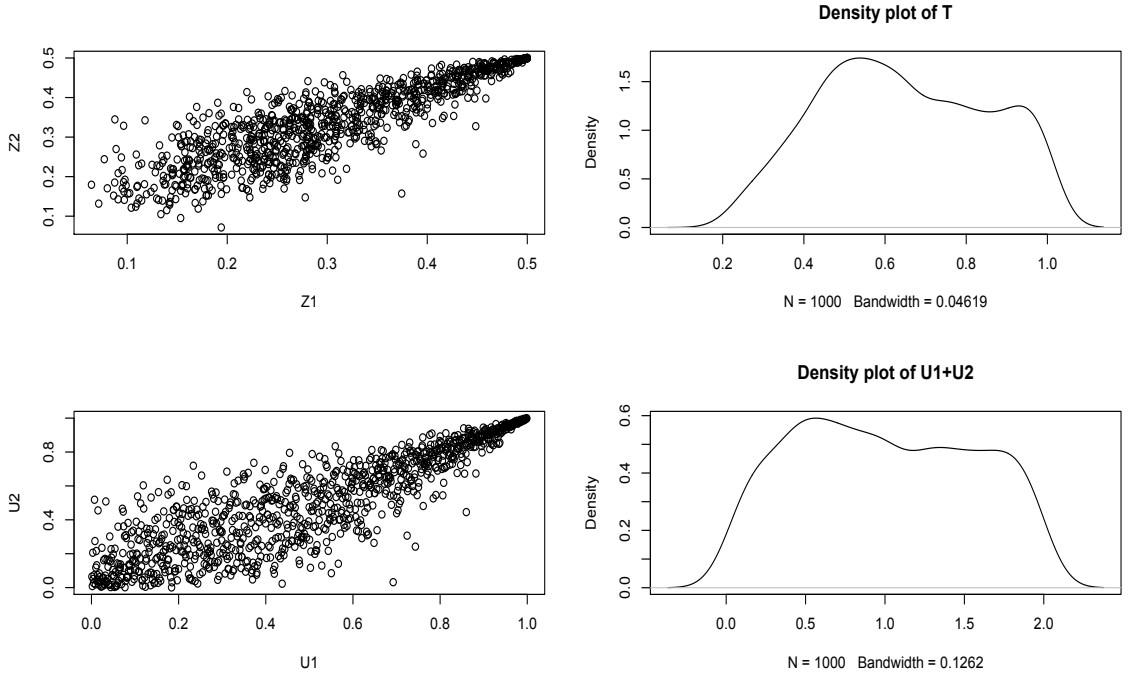


FIGURE G.9: Graphics of data set 6

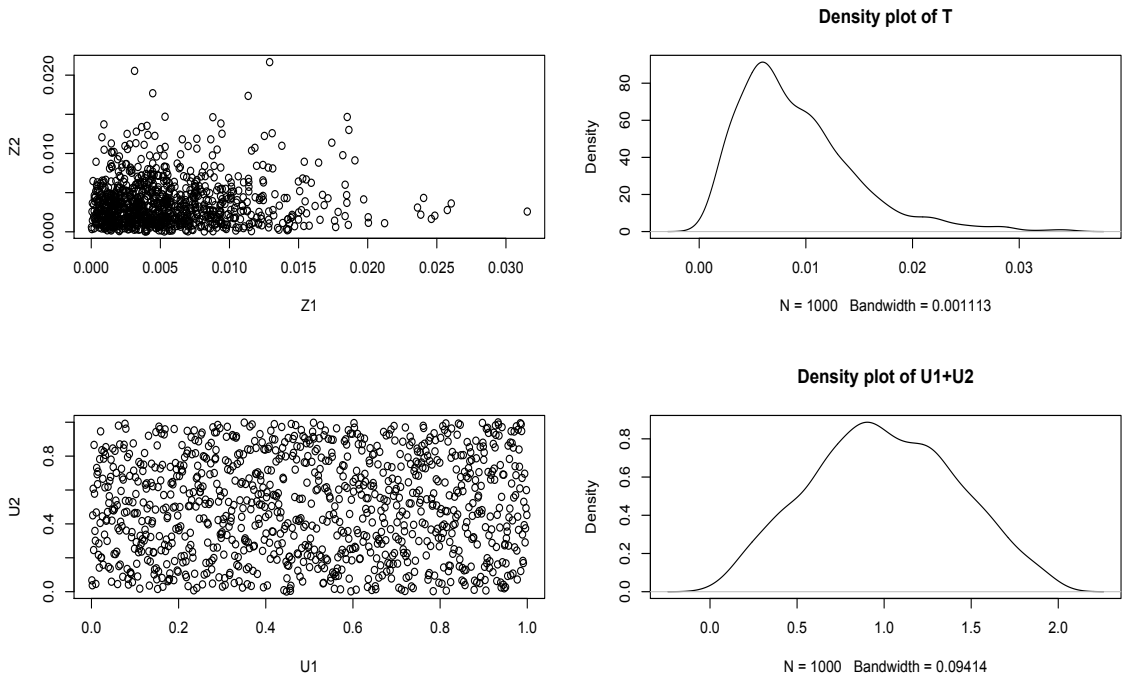


FIGURE G.10: Graphics of data set 7

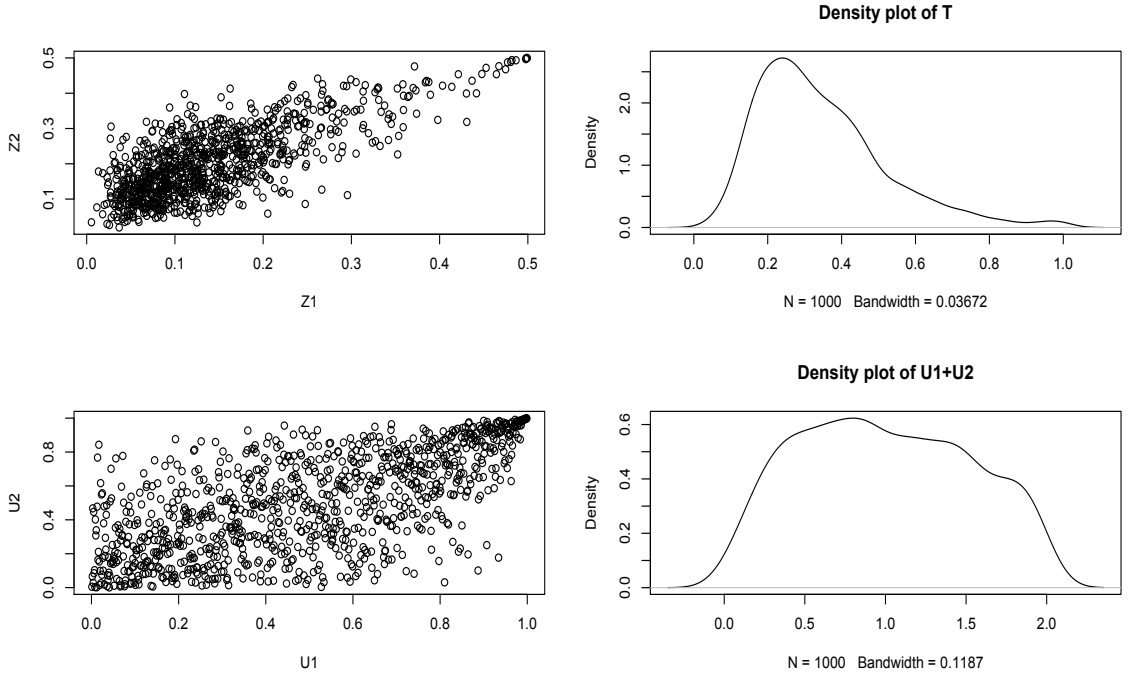


FIGURE G.11: Graphics of data set 8

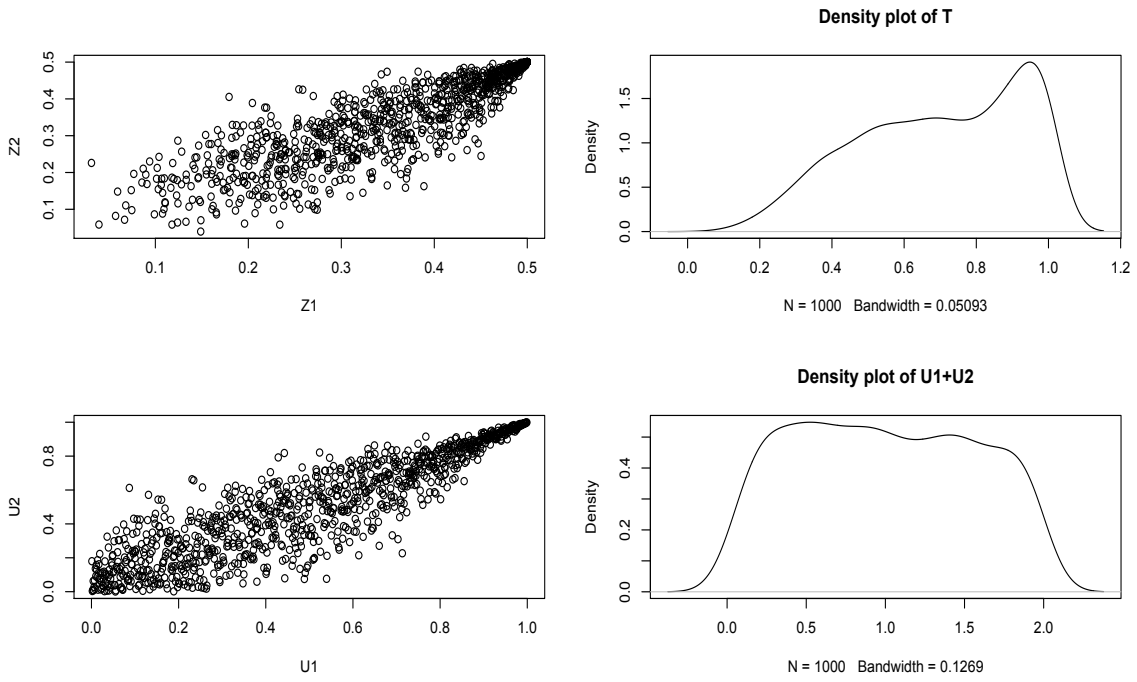


FIGURE G.12: Graphics of data set 9

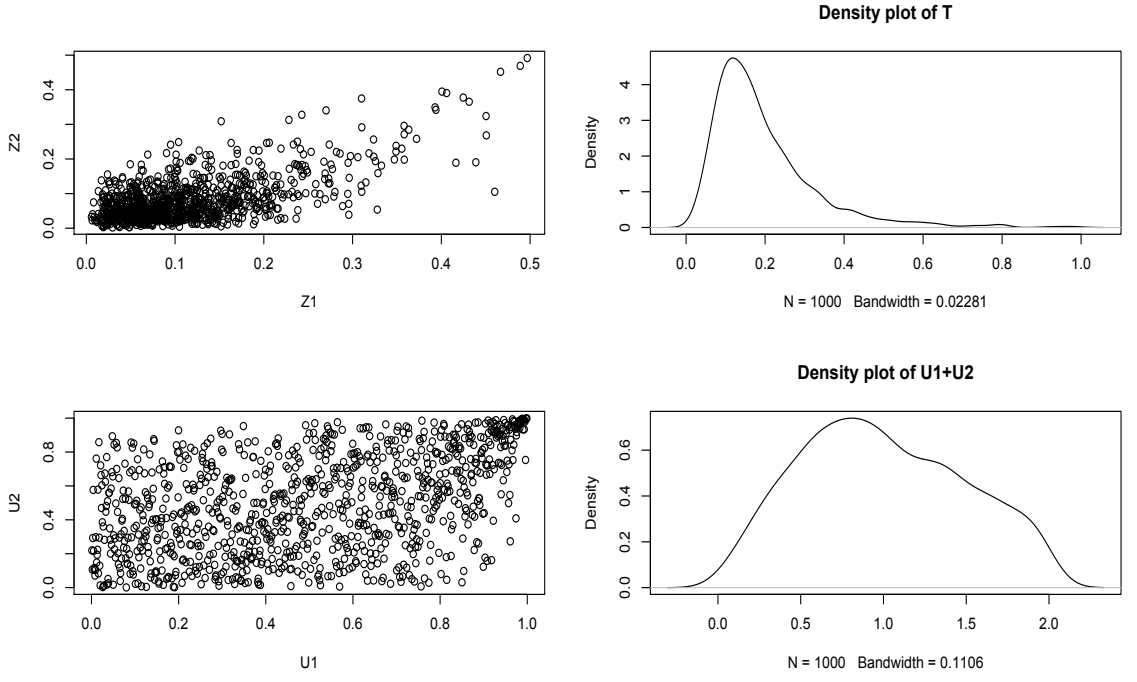


FIGURE G.13: Graphics of data set 10

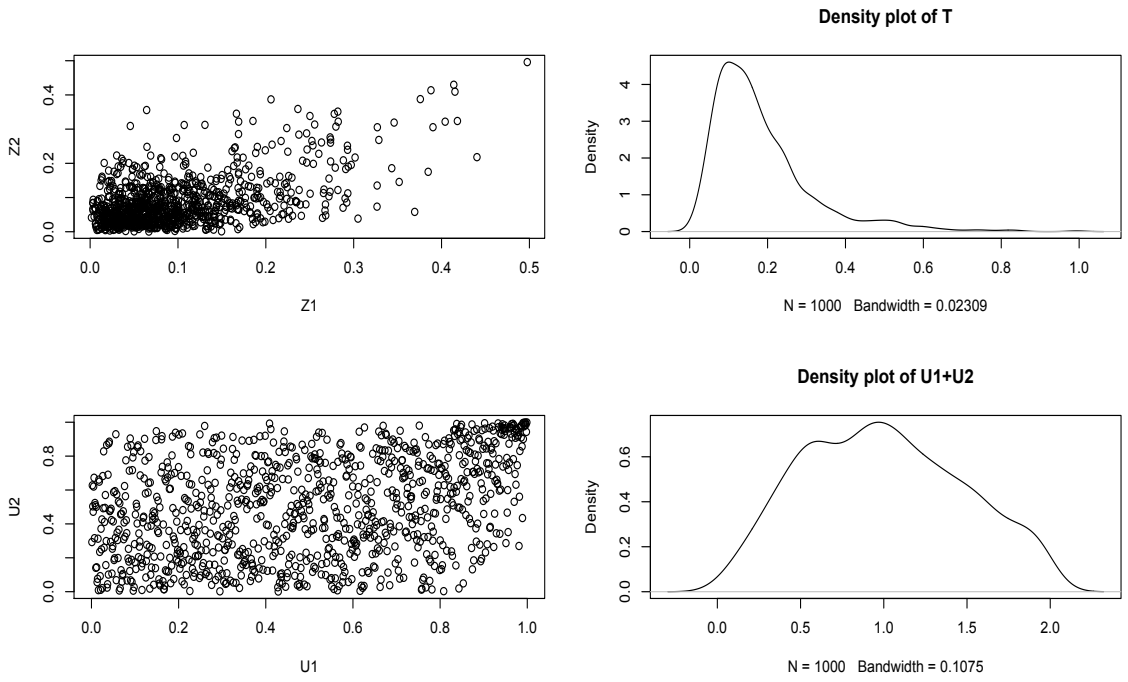


FIGURE G.14: Graphics of data set 11

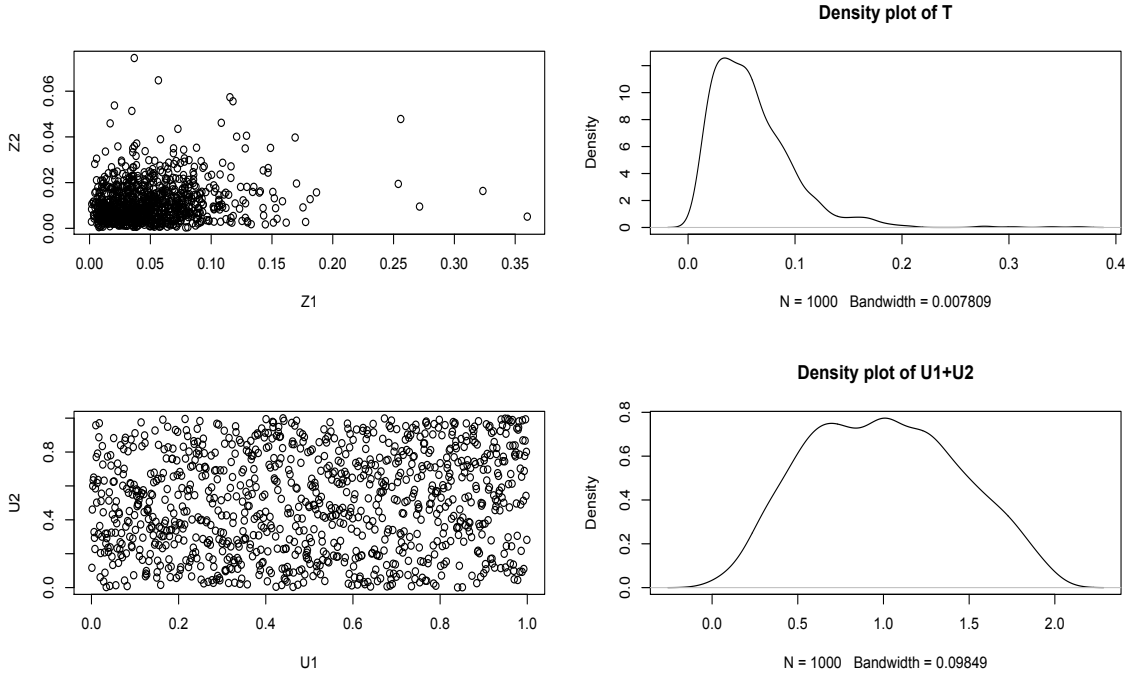


FIGURE G.15: Graphics of data set 12

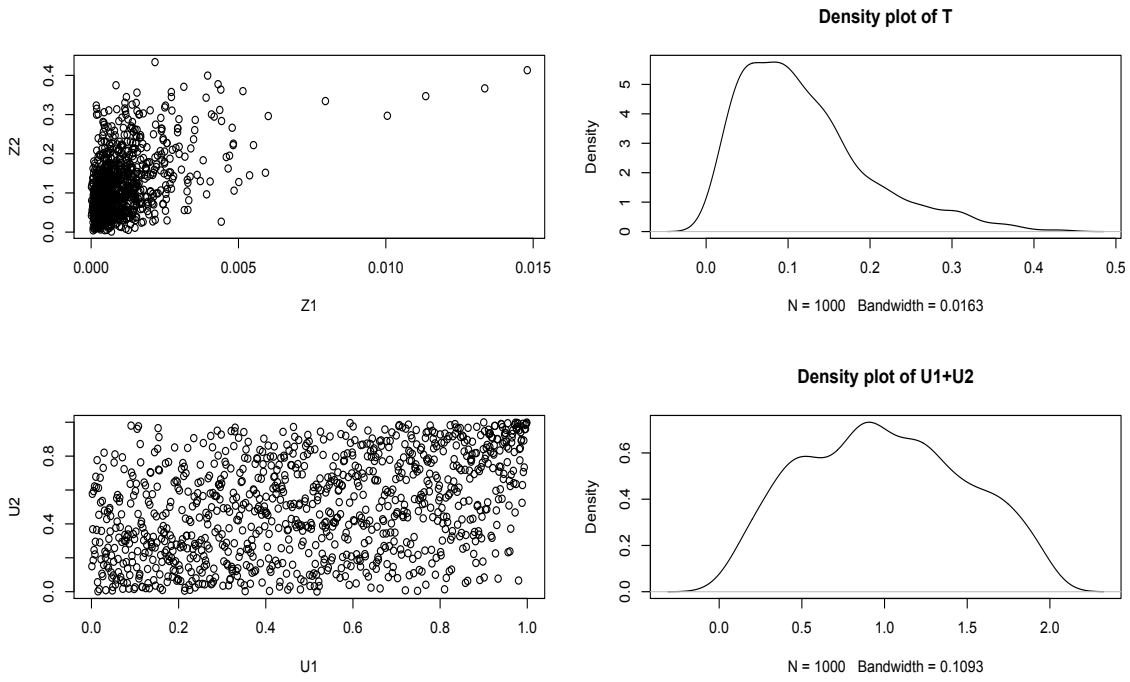


FIGURE G.16: Graphics of data set 13

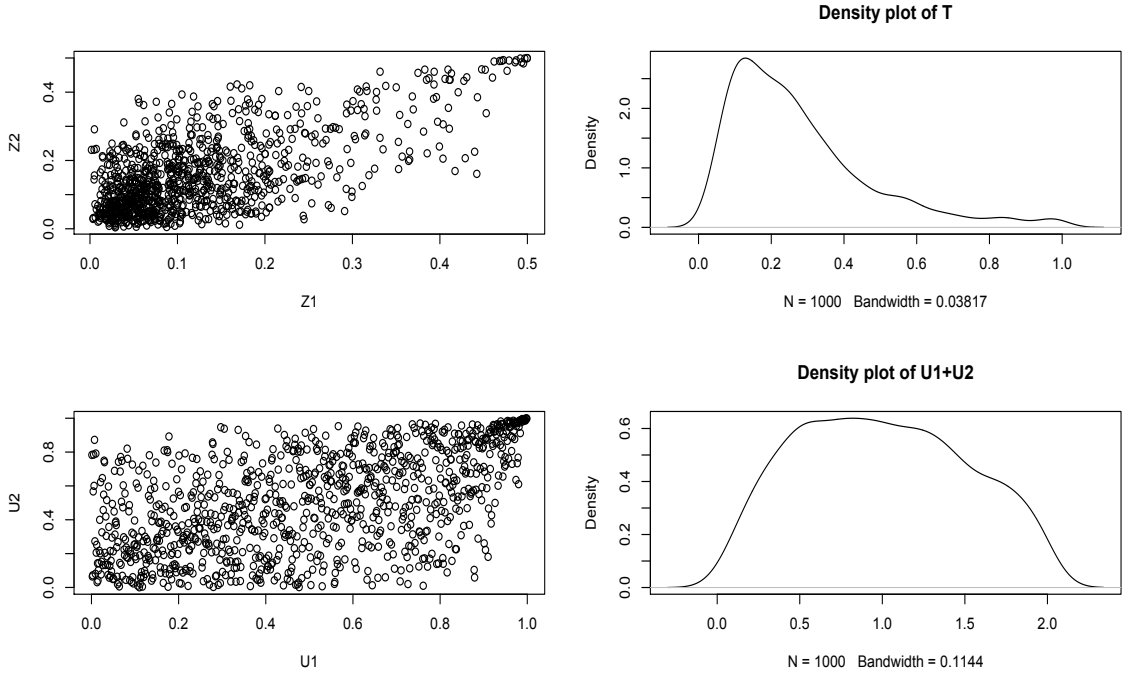


FIGURE G.17: Graphics of data set 14

Bibliography

[1] Jeffrey D. Allen, Min Chen, and Yang Xie. Model-Based Background Correction (MBCB): R Methods and GUI for Illumina Bead-array Data. *Journal of Cancer Science and Therapy*, 1(1):25–27, 2009.

[2] Dhammika Amaratunga and Javier Cabrera. *Exploration and analysis of DNA microarray and protein array data.* John Wiley and Sons, 2004.

[3] Jangsun Baek, Young Sook Son, and Geoffrey J. MacLachlan. Segmentation and intensity estimation of microarray images using a gamma-t mixture mode. *Bioinformatics*, 23(4):458–465, 2007.

[4] Benjamin M. Bolstad. *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.* PhD thesis, University of California, California, Berkeley, 2004.

[5] Benjamin M. Bolstad, Rafael A. Irizarry, M. Astrand, and Terence P. Speed. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics*, 19(2):185–193, 2003.

[6] Richard Bourgon, Robert Gentleman, and Wolfgang Huber. Independent filtering increases detection power for high-throughput experiments. *Proceeding National Academy of Sciences*, 107(2), 2010.

[7] Penelope A. Bryant, Gordon K. Smyth, and Roy Robins-Browneand Nigel Curtis. Technical variability is greater than biological variability in microarray experiment but both are outweighted by changes induced by stimulation. *Plos One*, 6(5):e19556, May 2011.

[8] Min Chen, Yang Xie, and Michael D. Story. An Exponential-Gamma Convolution Model for Background Correction of Illumina BeadArray Data. *Communication in Statistics: Theory and Methods*, 40(17):3055–3069, 2011.

[9] Sung E. Choe, Michael Boutros, Alan M. Michelson, George M. Church, and Marc S. Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(R16), 2005.

[10] Leslie M. Cope, Rafael A. Irizarry, Harris A. Jaffee, Zhijin Wu, and Terence P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20:323–331, 2004.

[11] Camilo Dagum. A New Model of Personal Income Distribution: Specification and Estimation. *Economie Appliquée*, 30(413 - 437), 1977.

[12] Liang-Hao Ding, Yang Xie, Seongmi Park, Guanghua Xiao, and Michael D. Story. Enhanced identification and biological validation of differential gene expression via Illumina whole-genome expression arrays through the use of the model-based background correction methodology. *Nucleic Acids Research*, 36(10: e58), 2008.

[13] Sørin Draghici. *Data analysis tools for DNA microarrays*. Chapman and Hall, 2003.

[14] Mark J. Dunning, Nuno L. Barbosa-Morais, Andy G. Lynch, Simon Tavaré, and Matthew E. Ritchie. Statistical issues in the analysis of Illumina data. *BMC Bioinformatics*, 9(85), 2008a.

[15] Rohmatul Fajriyah. Statistical analysis of the economic performance in Indonesia, Part I - Simplex method. 55th ISI Session Conference, April 2005a.

[16] Rohmatul Fajriyah. Statistical analysis of the economic performance in Indonesia, Part II - Grad method. ICREM 2 Conference, INSPEM, University Putra Malaysia, May 2005b.

[17] Rohmatul Fajriyah. The pdf's estimation by grad method and its Gini index. *Karya Asli Lorekan Ahli Matematik*, 1(2):021 – 027, 2008.

[18] Rohmatul Fajriyah. Comparison of the risk ratio of background correction models for the Illumina BeadArrays. Poster Paper EMS 2013, July 2013.

[19] Rohmatul Fajriyah. A study of convolution models for background correction of BeadArrays. *accepted paper at Austrian Journal of Statistics*, 2014a.

[20] Rohmatul Fajriyah. Generalized beta convolution model of the true intensity for the Illumina BeadArrays. *To appear at Thailand Statistician Association Journal*, 2014b.

[21] Rohmatul Fajriyah. The power and error rate of the special case of the cross variance test. *Submitted paper*, 2014e.

[22] Jian-Bing Fan, Sean X. Hu, William C. Craumer, and David L. Barker. Beadarray^{TM}-based solutions for enabling the promise of pharmacogenomics. *Bio Techniques*, 39:583–588, 2005.

[23] Jian-Bing Fan, Kevin L. Gunderson, Marina Bibikova, Joanne M. Yeakley, Jing Chen, Eliza Wickham Garcia, Lori L. Lebruska, Marc Laurent, Richard Shen, and David Barker. [3] illumina Universal Bead Arrays. *Methods in Enzymology*, 410:57–73, 2006.

[24] Anyiawung Chiara Forcheh, Geert Verbeke, Adetayo Kasim, Dan Lin, Ziv Shkedy, Willem Talloen, Hinrich WH Gohlmann, and Lieven Clement. Gene Filtering in the Analysis of Illumina Microarrays Experiments. *Statistical Applications in Genetics and Molecular Biology*, 11(2), 2012.

[25] D. A. Freedman. On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician*, 60:299 – 302, 2006.

[26] Magdalena Gabig and Grzegorz Wegrzyn. An introduction to DNA Chips: principles, technology, applications and analysis. *Acta Biochimica Polonica*, 48(3):615 – 622, 2001.

[27] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering*, 12(4):347–368, 2007.

[28] Christian Genest, Bruno Rémillard, and David Beaudoin. Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44:199–213, 2009.

[29] Michael B. Gordy. A Generalization of the Generalized Beta Distribution. Technical report, Board of Governors of the Federal Reserve System, Washington, 1998.

[30] Monique Graf and Desislava Nedyalkova. Fitting the Generalized Beta Distribution of the Second Kind to the Empirical Income Distribution from the Aggregate Laeken Indicators, February 2010. URL http://www.statistik.tuwien.ac.at/ameli/presentations/Fri1/GrafNedyalkova1.pdf.pdf.

[31] Monique Graf, Desislava Nedyalkova, Ralf Münnich, Jan Seger, and Stefan Zins. Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion. Technical Report 2.1, AMELI, 2011.

[32] Amber J Hackstadt and Ann M Hess. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, 10(11), 2009.

[33] Sepp Hochreiter, Djork-Arné, and Klaus Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949, 2006.

[34] Jorge D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature*, 7(200 – 210), 2006.

[35] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1: Statistics, pages 221 – 233, Berkeley ,California, 1967. Univ. of Calif. Press.

[36] Wolfgang Huber, Anja von Heydebreck, and Martin Vingron. Error models for microarray intensities. Technical Report Paper 6, Bioconductor Project Working Papers, 2004.

[37] Wolfgang Huber, Anja von Heydebreck, and Martin Vingron. An introduction to low-level analysis methods of DNA microarray data. Technical Report Paper 9, Bioconductor Project Working Papers, 2005a.

[38] Wolfgang Huber, Rafael A. Irizarry, and Robert Gentleman. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter Prepocessing Overview. Springer, 2005b.

[39] Rafael A. Irizarry and Zhijin Wu. *affycomp: Graphics Toolbox for Assessment of Affymetrix Expression Measures*, R package version 1.38.0 (with contributions from Simon Cawley) edition, 2013.

[40] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed. Summaries of Afymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4), 2003a.

[41] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4(2):249–264, 2003b.

[42] Rafael A. Irizarry, Zhijin Wu, and Harris A. Jaffee. Comparison of Affymetrix geneChip expression measures. *Bioinformatics*, 22(7):789–794, 2006.

[43] Mei-Ling Lee. *Analysis of microarray gene expression data*. Springer, New York, 2006.

[44] Lawrence M. Leemis and Jacquelyn T. McQueston. Univariate Distribution Relationships. *The American Statistician*, 62(1):45–53, February 2008.

[45] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceeding national Academy of Sciences*, 98(1):31–36, 2001.

[46] James B. McDonald. Some generalized functions for the distribution of income. *Econometrica*, 52(3), May 1984.

[47] James B. McDonald and Yexiao J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66:133–152, 1995.

[48] Monnie McGee and Zhongxue Chen. Parameter estimation for the convolution model for background correction of affymetrix genechip data. *Statistical Applications in Genetics and Molecular Biology*, 5(24), 2006.

[49] Monnie McGee, Zhongxue Chen, Richard H. Scheuermann, and Feng Luo. A nonparametric background correction method for oligonucleotide arrays. Technical Report TR340, Southern Methodist University, 2006.

[50] Geoffrey J. McLachlan, Kim-Anh Do, and Christopher Ambroise. *Analyzing Microarray Gene Expression Data*. John Wiley and Sons, New Jersey, 2004.

[51] Melissa B. Miller and Yi-Wei Tang. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. *Clinical Microbiology Reviews*, 22(4):611–633, 2009.

[52] D. Pfeffermann, C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, 60(Part 1):23 – 40, 1998.

[53] Sandra Plancade, Yves Rozenholc, and Eiliv Lund. Improving background correction for Illumina BeadArrays: the normal-gamma model, 2011.

[54] Sandra Plancade, Yves Rozenholc, and Eiliv Lund. Generalization of the normal-exponential model: exploration of a more accurate parameterisation for the signal distribution on Illumina BeadArrays. *BMC Bioinformatics*, 13 (329), 2012.

[55] Alexandra Posekany, K. Felsenstein, and Peter Sykacek. Biological assessment of robust noise models in microarray data analysis. *Bioinformatics*, 27(6): 807–814, 2011.

[56] Serge Provost and Edmund M. Rudiuk. The exact density function of the ratio of two dependent linear combinations of chi-square variables. *Annals of the Institute of Statistical Mathematics*, 46(3):557–571, 1994.

[57] Aladdin Shamilov, Yeliz Mert Kantar, and Ilhan Usta. On a Functional defined by means of Kullback-Leibler Measure and Its Statistical Applications. In *Proceedings of the 9th WSEAS International Conference on Applied Mathematics*, pages 632–637, May 2006.

[58] Wei Shi, Alicia Oshlack, and Gordon K. Smyth. Optimizing the noise versus bias trade-off for Illumina whole genome expression Beadchips. *Nucleic Acids Research*, 38(22: e204), 2010.

[59] Jeremy D. Silver, Matthew E. Ritchie, and Gordon K. Smyth. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution model. *Biostatistics*, 10:352–363, 2009.

[60] Gordon K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[61] Frank J. Steemers and Kevin L. Gunderson. Illumina, Inc. *Pharmacogenomics*, 6:777–782, 2005.

[62] Timothy J. Triche, Daniel J. Weisenberger, David Van Den Berg, Peter W. Laird, and Kimberly D. Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, pages 1–11, March 2013.

[63] Maarten van Iterson, Judith M Boer, and Renée X Menezes. Filtering, FDR and power. *BMC Bioinformatics*, 11(450), 2010.

[64] Anja von Heydebreck, Wolfgang Huber, and Robert Gentleman. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, chapter Differential expression with the Bioconductor project. John Wiley and Sons, 2004.

[65] Levi Waldron. *ffpe: Quality assessment and control for FFPR microarray expression data*, r package version 1.4.0 edition, 2013.

[66] Levi Waldron, Shuji Ogino, Yujin Hoshida, Kaori Shima, Amy E. McCart Reed, Peter T. Simpson, Yoshifumi Baba, Katsuhiko Nosho, Nicola Segata, Ana Cristina Vargas, Margaret Cummings, Sunil R. Lakhani, Gregory J. Kirkner, Edward Giovannucci, John Quackenbush, Todd R. Golub, Charles S. Fuchs, Giovanni Parmigiani, and Curtis Huttenhower. Expression Profiling of Archival Tumors for Long-term Health Studies. *Clinical Cancer Research*, 2012.

[67] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez-Murillo, and Forrest Spencer. A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, 99(468):909 – 917, 2004.

[68] Yang Xie, Xinlei Wang, and Michael D. Story. Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*, 25(6):751–757, 2009.

[69] Aidong Zhang. *Advanced analysis of gene expression microarray data*. World Scientific Publixhing, Singapore, 2006.

[70] Jin Zhang. Reducing the bias of the maximum likelihood estimator of the shape parameter for the gamma Distribution. *Computational Statistics*, 28: 1715 – 1724, 2013.

[71] Wen Zhu, Nancy Zeng, and Ning Wang. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. http://www.nesug.org/Proceedings/nesug10/hl/hl07.pdf, 2010.