GENE EXPRESSION INDICES FOR SINGLE-CHANNEL MICROARRAYS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TÜLAY AKAL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

APRIL 2013

Approval of the thesis:

GENE EXPRESSION INDICES FOR SINGLE-CHANNEL MICROARRAYS

submitted by **TÜLAY AKAL** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İnci Batmaz _____
Head of Department, **Statistics**

Assoc. Prof. Dr. Vilda Purutçuoğlu _____
Supervisor, **Statistics Dept.**

Examining Committee Members:

Prof. Dr. Gerhard Wilhelm Weber
Applied Mathematics Dept., METU _____

Assoc. Prof. Dr. Vilda Purutçuoğlu
Statistics Dept., METU _____

Prof. Dr. Ayşen Dener Akkaya
Statistics Dept., METU _____

Prof. Dr. İnci Batmaz
Statistics Dept., METU _____

Assist. Prof. Dr. Berna Burçak Başbuğ Erkan
Statistics Dept., METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: TÜLAY AKAL

Signature :

ABSTRACT

GENE EXPRESSION INDICES FOR SINGLE-CHANNEL MICROARRAYS

Akal, Tülay

M.S., Department of Statistics

Supervisor : Assoc. Prof. Dr. Vilda Purutçuoğlu

April 2013, 122 pages

The microarray technology is one of the recent and advance tools in biological sciences. This optical technology aims to measure the amount of changes in transcripted message for each gene by RNA via quantifying the colour intensity on the arrays. But due to the different experimental conditions, these measurements can include both systematic and random erroneous intensities.

In this study, we deal with one of these systematic sources of errors, called background signals, for one-channel microarrays. Hereby, we initially describe the most well-know methods such as MAS 5.0, MBEI, RMA, and BGX approaches for estimating the gene expression levels, i.e., gene expression indices. Then, we present a novel gene expression index, called multi-RGX (Multiple Probe-Robust Gene Expression Index), which can be seen as a generalization of the FGX model and closely related to the BGX method developed for this type of arrays. In multi-RGX, the FGX model is extended by both covering nonnormal log-expressions, in particular, long-tailed symmetric (LTS) densities, and taking not only the probe mean intensities, rather using all gene expressions in each probe for every gene. In inference of such model, we apply the modified maximum likelihood method to deal with the unexplicit solutions of the likelihood equations under LTS. Moreover, we derive the covariance-variance matrix of model parameters from the observed Fisher Information matrix. Finally in order to find the gain in information from the estimation, we evaluate the performance of our novel index in different datasets.

Keywords: Background normalization, microarray, oligonucleotide, modified maximum like-

lihood estimators, observed Fisher information matrix

ÖZ

TEK-KANALLI MİKRODİZİNLERİN GEN İFADE İNDEKSLERİ

Akal, Tülay
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi : Doç. Dr. Vilda Purutçuoğlu

Nisan 2013, 122 sayfa

Mikrodizin teknolojisi biyoloji bilimlerindeki yeni ve gelişmiş araçlardan birisidir. Bu optik teknoloji, RNA tarafından her bir gen için transkrip edilen mesajdaki değişim miktarını, dizinlerdeki renk yoğunluğunu bularak ölçmeyi amaçlamaktadır. Fakat farklı deneysel koşullardan dolayı, bu ölçümler hem sistematik hem de rassal hatalı yoğunlukları da içerebilir.

Bu çalışmada, tek - koşullu mikrodizinlerde, sistematik hata kaynaklarından biri olan ardalan sinyalleriyle ilgilenmekteyiz. Bu amaçla, öncelikle gen ifade düzeylerini, yani gen ifade indekslerini, tahmin etmek için kullanılan, en çok bilinen, MAS 5.0, MBEI, RMA, ve BGX yaklaşımları gibi yöntemleri tanıtmaktayız. Daha sonra çoklu-RGX (Çoklu Prob-Sağlam Gen İfade İndeksi) adlı, bu çeşit dizinler için geliştirilmiş, FGX modelinin genelleştirilmiş hali olarak görülebilen ve BGX yöntemiyle de oldukça bağlantılı olan, yeni bir gen ifade indeksi sunmaktayız. Çoklu-RGX'de, hem normal olmayan, özellikle uzun kuyruklu simektrik (LTS) dağılımlı, logaritmik ifadeleri kapsayarak, hem de FGX modelin uyguladığı gibi sadece prob ortalama yoğunluklarını almak yerine, her gen ve her bir probdaki tüm gen ifadelerini kullanarak FGX modeli genişletilmiştir. Böyle bir modelin tahmininde, LTS altında olabilirlik denklemlerinin açık olmayan sonuçlarını çözebilen, uyarlanmış en çok olabilirlik yöntemini uygulamaktayız. Ayrıca gözlemlenebilir Fisher Bilgi Matrisi yardımıyla, model parametrelerinin kovaryans-varyans matrisini çıkarmaktayız. Son olarak, ölçümlerden gelen bilgi artışını bulmak için yeni indeksimizin FGX'e göre başarısını farklı veri kümeleriyle değerlendirmekteyiz.

Anahtar Kelimeler: Ardalan normalizasyonu, mikrodizin, oligonükleotid, uyarlanmış en yüksek

olabilirlik tahminleyicileri, gözlemlenen Fisher bilgi matrisi

To my lovely family

Arslan Akal, Sevim Akal, Orhan Akal, Nilgün Akal, Büşra Akal, Edanur Akal

ACKNOWLEDGMENTS

This thesis reinforced my patience and my interest in academic study and advanced them to an upper level. In fact, my Master's study contributed to this process very much. Firstly, I want to express my gratitude to my precious and lovely supervisor, Assoc. Prof. Dr. Vilda Purutçuoğlu, for all of her efforts and supports, and being much more than a supervisor to me. Throughout writing the thesis, she has provided suggestions and guided all the processes. We have been working since 2010 and I am very thankful that she has given me an opportunity to work with her. She has a significant role in my academic life and I will follow her in my life.

I want to thank many people who have contributed to my academic life, and the most significant ones are my lovely family members, my aunts, uncles, grandmother and grandfathers. I am really grateful for all their efforts, for being a real family to me and supporting me under any condition. Also, I am grateful for all the supports that my dear departed grandfather İbrahim Akal provided me. I hope he rests in peace.

I have really appreciated help of my dear friends: Tuğba Yılmaz, Ferin Merve Yılmaz, Seçil Çaşkurlu, Zeynep Arslan, Fatma Işkın, Yeşim Türk, Fatma Irk, Meliha Sermin Paksoy and Habibe Saka, for all the supports they given to me throughout my thesis period.

Also, I would like to take this opportunity to thank a friend of me, for his guidance during this study. He really showed me a different window to look through.

Moreover, I want to thank all of my teachers, lecturers, professors from primary school to university.

I would like to thank to my dear professors being in my thesis' jury and their valuable feedbacks and supports, which they provide.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

The *genome*, which represents all the genetic information of a living organism, has always been in concern for scientists. This structure is composed of long deoxyribonucleic acid, also shortly called *DNA*, molecules. These molecules are known as *chromosomes*. Each cell of an organism consists of DNA in their nucleus and the *gene* is a specific part of DNA. As it can be seen in Figure 1.1, it has a double helical structure, in other words, it is *double stranded* that is basically generated with two-polynucleotide chains bonded with hydrogen bonds. DNA is a polymer of a number of nucleotides containing a sugar, base and a phosphate. In DNA, a nucleotide consists of four types of bases, which are *adenine* (A), *thymine* (T), *guanine* (G) and *cytosine* (C) (Grant et al., 2013). Here adenine binds with thymine (A-T) and guanine binds with cytosine (G-C).

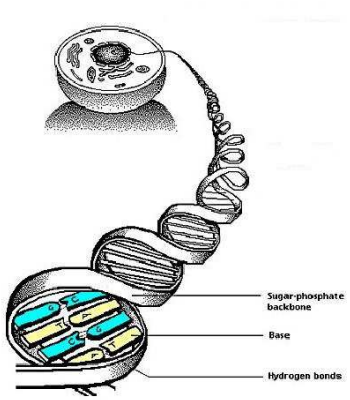


Figure 1.1: Double helical structure of DNA (modified from Grant et al. (2013)).

In the procedure of the protein synthesis, the synthesis of another molecule, called the *Ribonucleic acid* (RNA) is required. Unlike DNA, RNA is usually *single stranded* and it has the base uracil, instead of thymine (Grant et al., 2013). While DNA has deoxyribose as sugar, RNA contains ribose.

To synthesize a specific protein, firstly, DNA is copied into a type of RNA, called the *messenger RNA* (mRNA), being a complementary to its DNA template. This process is called the *transcription*. Then another type of RNA, which transfers RNA (tRNA) brings the amino acids to the organelle, called *ribosome*, which is responsible from protein synthesis in a cell. At last, the process of the protein synthesis is carried out. This process is known as the *trans-*

lation (Grant et al., 2013). The transcription and the translation are also named as the *central dogma*. The process of the very basic protein synthesis is shown in Figure 1.2.

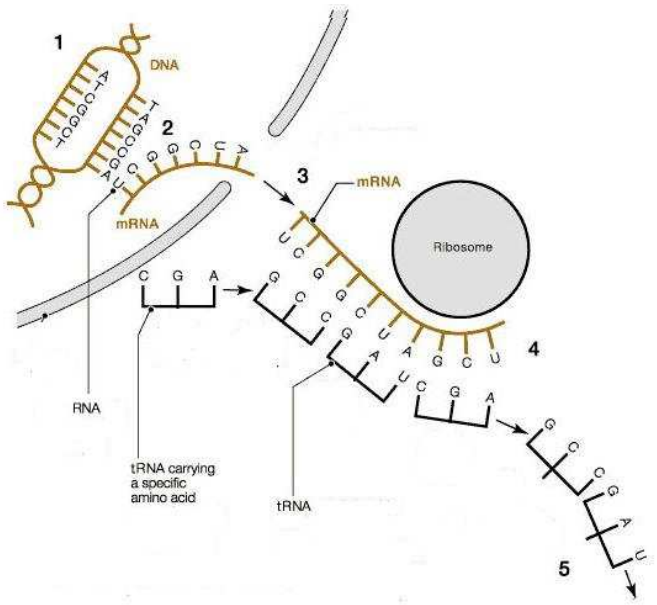


Figure 1.2: Simple representation of the protein synthesis (modified from Grant et al. (2013)).

The synthesized protein molecule is very important in the sense that it gives information about the cell with respect to the active genes. Furthermore, the gene expression is gathered while transferring from DNA to RNA and from RNA to the synthesis of protein molecules (Grant et al., 2013). Thus, it is a significant phenomenon in order to understand the biological organisms. Accordingly, the microarray technology gives opportunity in this manner (Kerr, 2009) in the sense that it enables us to analyze the behavior of genes under different conditions (Sanchez and Ruiz de Villa, 2008).

In the scope of the microarray analysis, there are some companies, such as Aligent and Affymetrix, providing microarray platforms. Those make use of specificity and affinity of complementary base-pairing of nucleic acid. A *microarray*, also known as the *DNA chip*, *gene chip* or *biochip*, consists of thousands of DNA sequences, called *probes*, which are attached to a solid surface. In order to measure the gene expressions, the extracted genetic material, called the *target*, is labeled with a florescent dye. After hybridization in the probes, using a special scanner, the amount of florescence is measured, and it is called the *intensity* (Kerr, 2009). Hereby, the amount of mRNA transcripts is measured and it can be seen as an approximation to the level of expression of the gene and the measurements are turned into pictures (Sanchez and M. C. R. de Villa, 2008). A layout of microarray experiment can be seen in Figure 1.3.

The oligonucleotide is a type of single channel microarray and the Affymetrix GeneChip is an example of the popular oligonucleotides.

In these arrays, also called chips, we observe probes, which are the known segments of particular gene sequences. Each probe contains two parts, namely, the *perfect match* (PM) and the *mismatch* (MM). The former stands for the perfect transcription of the cRNA and the latter is aimed to measure the faulty signals on the arrays by changing the 13th base pair of the PM.

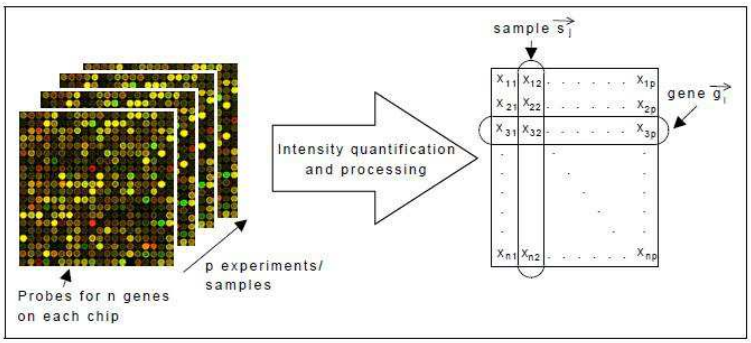


Figure 1.3: Simple representation of a microarray experiment (modified from Kerr (2009)).

In a microarray study, we can define three main sources of variation of signals, which are also represented as the systematic error. The first one is named as the *nonspecific hybridization*, which is caused by the misbinding of some parts of the target sequence to non-complementary transcripts. This error measures the binding, which leads to the greater intensity value than the actual measurement on the arrays. On the other hand, the second source of variation is called the *background signal* that is the one gathered under no hybridization, hereby free from any true signal. Finally, the third source of noisy signal is known as the *stray signal*, which may be generated by binding the sequence to the surface of the slide, rather than to the probe, resulting in variations in the intensity levels. We can call these three types of variations as the *systematic error* (Purutçuoğlu, 2007; Purutçuoğlu, 2012).

On the other side, the term *gene expression index* stands for the mathematical method to estimate the true expression level in the oligonucleotides.

There are a number of indices developed for this purpose. In this study we particularly deal with the most well-known and current indices such as MAS 5.0, RMA, dChip (or MBEI), GC-RMA, BGX, FGX, and RGX by specifying where they are applied in the biological literature. Each of them has their own advantages and disadvantages.

For instance, MAS 5.0 includes bias in inference of the true signal since it is based on the ad-hoc adjustment of the intensity when PM values are measured less than the associated MM probes. On the other hand, the MBEI method can infer the true signals without ad-hoc calculations, whereas, it cannot make estimation for large number of arrays as the inference is done via the least square method. Similarly, RMA also fails under this condition. But it is more sensitive than MBEI in terms of the detection of differential expressed genes. On the other hand, GC-RMA gives more accurate results than MAS 5.0 and RMA. But it completely ignores the information from MM in the estimation of the signals. FGX is advantageous over its alternatives for reducing bias and computational demand. However it is based on a strict normality assumption. Whereas, RGX can deal with non-normal densities and it is computationally as fast as FGX. Moreover, we highlight the most recent studies based on microarray normalization in general. For this purpose, we consider the doctorate thesis of

Ülgen (2010) and discuss the similarity and distinction with this thesis.

Hereby in this study, we initially explain the underlying most well known gene expression indices and some recent methods in details and describe the current idea of normalization in general. Then, we present our novel gene expression index, called multi-RGX, in order to solve the problem of recently developed methods, which are FGX and RGX, and emphasize the differences between the current researches. In this new method, we consider gene and probe specific signal in the measurement of microarray and develop explicit expressions for each model parameter via the modified maximum likelihood method. Furthermore, as the second novelty, we present the explicit forms of the variances and covariances of model parameters via the Fisher Information Matrix. Moreover, we also represent other alternative choices of the multi-RGX model and state the estimators of model parameters. As given in the associated chapter, those alternative approaches do not produce explicit formula, hereby, are still iterative procedures. Furthermore, as one of the major aims of our study is to suggest computationally fast and accurate method, we evaluate those alternatives gene expression indices with our novel method in the application parts.

Accordingly, in the thesis, we explain the idea of the modified maximum likelihood estimators and assess the performance of multi-RGX with its strong competitive with respect to different criteria such as the signal detect R^2, R^2, signal detect slope and CPU (Central Processing Unit) time. Each of these criteria is presented in Chapter 4 and the code of the function, which is originally developed in this study, is given in Appendix. In the assessment, we use four different datasets. The first two data are benchmark datasets from Affymetrix and GeneChips brands, respectively. For those sets, we compare all the well-known and current indices via distinct criteria. On the other hand, as the third and fourth datasets, we use simulated measurements and compare only the results of FGX, RGX and our novel index, multi-RGZX, due to the fact that these are the strong alternatives of multi-RGX and this index is indeed developed to overcome the challenges of these two models. Finally, we report our results and discuss our future directions in Conclusion.

As a result, we organize the thesis as the following plan. The recent studies and alternative approaches in background normalization are presented in Chapter 2. Chapter 3 is dedicated to our new algorithm (multi-RGX), mathematical derivation of model parameters and their covariance-variance terms. Here, we also represent plausible alternative modelling of multi-RGX and declare why we choose our model with respect to others. In Chapter 4, we evaluate all methods via real and simulated datasets under different model selection criteria. Chapter 5 summarizes all outputs and suggests our future perspectives.

CHAPTER 2

BACKGROUND NORMALIZATIONS' METHODS

As stated beforehand, there are a number of gene indices in the literature, which aim to measure the true signals of the gene expression from noisy data. These indices are under the background normalization approach, which is one of the steps in the overall normalization of the microarray data.

There are two sources of error in the observations. These are randomized and systematic errors. The *normalization* is the process to discard the systematic errors in the observed microarray dataset. This process is typically implemented before the analysis of the data. Therefore, it can be considered as the *preprocessing steps* of the actual analysis. The randomized error cannot be discarded and its presence does not cause any bias in the analysis. Whereas, the systematic error implies other sources of variations, which are not originated from the changes in the gene expressions. Therefore, it may lead to bias in the analysis if it is not eliminated from the measurements (Wit and McClure, 2004; Steen, 2002; Stekel, 2003).

The possible sources of the systematic errors can be separated under the three groups for one-channel microarrays. These are

1) *Spatial normalization*, which enables us to exclude errorenous signals due to the problems during the scanning of the array, unevenly washing the chips or localization of the array (Wit and McClure, 2004; Steen, 2002; Stekel, 2003).

2) *Background normalization*, which can detect any errorenous signal due to the non-specific hybridization faulty signals in the probe or array on the scanners (Wit and McClure, 2004; Steen, 2002; Stekel, 2003).

3) *Within-between array normalization*, which can handle any possible signals due to the design of the genes/conditions on the arrays (Wit and McClure, 2004; Steen, 2002; Stekel, 2003).

On the other hand, an alternative approach suggested by Kerr et al. (2000) and Kerr and Churchill (2001), the estimation of the signal and the analysis of the differentially expressed genes can be done within an ANOVA model in which the normalization is not applied separately as the preprocessing of the data before the actual analysis, rather, it can be done simultaneously within an ANOVA model. This idea is discussed and used in the analysis to capture the changes in gene expressions as well in the study of Ülgen (2010). In this model it is assumed that the gene expression on the ith array, jth probe, kth variety, i.e. condition, and the gth gene, denoted by y_{ijkg}, can be described on the logarithmic scale via

$$\log(y_{ijkg}) = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \tag{2.1}$$

in which μ is the average signal, A_i and V_k denote the ith array effect and the kth treatment/condition effect, respectively. Moreover, G_g shows the gth gene effect. Finally, $(AG)_{ig}$ and $(VG)_{kg}$ represent the interaction effect of array on the gth gene as well as treatment on the gth gene, in order. In the end, ε_{ijkg} stands for the random error, which comes from independent and identically distributed density with mean zero. In general, ε_{ijkl} is accepted to have a normal distribution (Kerr et al., 2000). But from the comparative analysis via real and simulated datasets under Dixon's outlier, mixture and contamination models, it is showed that the choice of long-tailed symmetric distribution for ε, which enables us to implement robust analysis, is more realistic than the strick normality assumption (Ülgen, 2010). In this study, the measurements are modelled under an unbalanced two-way classification fixed effect model with interaction such that

$$\log(y_{ijkg}) = \mu + A_i + V_k + G_g + (AG)_{ig} + (VG)_{kg} + \varepsilon_{ijkg} \tag{2.2}$$

where l indices the measurement from 1 to n_k , i.e., $1 \leq l \leq n_k$, and n_k denotes the number of observations in the kth treatment for every gene. In the estimation of the model parameters, as ε_{kgl} is not dependent on the strict normality, LTS density is accepted, resulting in estimators via the modified maximum likelihood (MML) method, which we use in this study as well and present its mathematical details in Chapter 3, and adaptive maximum likelihood (AMML) approach (Dönmez, 2010; Tiku and Sürücü, 2009), that is also known as revised MML or MML30. Moreover, the significance of the estimates in ANOVA model is evaluated via different pairwise multiple comparison testing procedure based on MML estimators and AMML, which are originally developed from the Dunnett (1982) pairwise t-test and simulated comparison test based on noncentral F and W statistics (Ülgen, 2010). The performance of estimates is then compared with least squares and Huber's M-estimators, respectively, regarding their powers and relative efficiencies.

On the other hand, from the previous analyses, it is shown that the ANOVA approach with respect to preprocessing procedure of normalization is computationally intensive, in particular, when more complicated normalizations are required (Wit and McClure, 2004). In this thesis, we accept that the researcher follows the second strategy that is based on the preprocessing calculation of the data in advance of the actual analysis. Accordingly, we initially present well-known and current background normalization methods before describing our novel approach as a plausible alternative of them. Then, we describe our suggested method in details.

2.1 MAS 5.0 Method

MAS 5.0 (*Microarray Suite Software*) method is one of the well known gene expression indices specifically developed for oligonucleotides. It is already used in a number of microarray analysis. For instance, it is implemented in a study to detect the role MAP kinase types in adult mouse hearts (Mitchell et. al., 2005), to identify, diagnose and predict the survival of

a lymphoma as well as lymphoproliferative disorder (Staudt et. al., 2011). Moreover, Klienstein et al. (2006) use it in the analysis about the diversity in Arabidopsis thaliana regarding the gene expressions. Reppe et al. (2007) perform it in a study to seek for the effect of abnormal muscles and hematopoieyic gene expressions on clinical morbidity in the primary hyperparathyroidism. Kostek et al. (2007) apply this method to determine the molecular mechanisms of lengthening and shortening constructions in human muscles. Additionally, Yang et al. (2012) conduct a study regarding the survival in head and neck cancers and perform this normalization approach and finally, Venezia et al. (2004) investigate the molecular signatures of proliferation and quiescence in hematopoietic stem cells via the microarray study with MAS 5.0 technique.

In this method, the true signal in the PM probe is considered to be affected by the stray signal in an additive way, and the stray signal is the unique source of the MM probe (Hubbell et al., 2002). In this approach, the true signal T is calculated as follows:

$$\log \mathrm{T}_{ij} = \log (\mathrm{PM}_{ij} - S_{ij}),$$

where $i = 1, 2, ..., n$, $j = 1, 2, ..., m$ and $T_{ij} = \mathrm{PM}_{ij}/S_{ij}$.

Here, S_{ij} denotes the stray signal and PM_{ij} describes the perfect match for the ith gene and the jth probe. Accordingly, i and j stand for the gene and probe indicator, respectively. In MAS 5.0, the decision of the amount of stray signals is presented as the following criterion:

- If $\mathrm{PM}_{ij} > \mathrm{MM}_{ij}$, then $S_{ij} = \mathrm{MM}_{ij}$, that is, the stray signal is thought to merely come from the mismatch.
- If $\mathrm{PM}_{ij} \leq \mathrm{MM}_{ij}$, then $\log \mathrm{S}_{ij} = \log \mathrm{PM}_{ij} - \mathrm{SB}_i^+$,

where S_{ij} represents the stray signal and PM_{ij} shows the perfect match. Moreover, SB_i^+ indicates a specific background and aims to give a robust estimate of the typical log intensity for gene i. Hence, it is calculated by

$$\mathrm{SB}_i^+ = T_{bi}[\log \mathrm{PM}_{ij} - \log \mathrm{MM}_{ij}]$$

in which T_{bi} presents the one-step Tukey biweight estimator of location and reports the information about the typical probe pair for the ith gene.

The Tukey biweight, also called the *bisquare weight*, is a robust statistic in the sense that it assigns the weights to data points $x_j (j = 1, 2, \ldots, n)$, which are expressed as $x_j = \log \mathrm{PM}_{ij} - \log \mathrm{MM}_{ij}$, with respect to their distances to the median. By this way, it aims to find a robust average, which is not affected from the outliers (Affymetrix, 2002). Thereby, it is computed as:

$$T_{bi} = \frac{x_j - \tilde{\mu}_i^m}{\tilde{\sigma}_i^m},$$

where $\tilde{\mu}_i^m$ is the median of the ith gene in the m probe, determining the center of the corresponding data and $\tilde{\sigma}_i^m$ denotes the median absolute deviation. The weight w is used for determining how much each value should contribute to the average and gives the advantage of reducing the effect of the outliers on the average. Hereby, it is found by

$$w(u) = (1 - u^2)^2, \quad \text{for } 0 \leq |u| \leq 1 \tag{2.3}$$

and zero otherwise. As seen in Equation (2.3), the outliers can be handled by a smooth function since the weights are reduced to 0 for those, which are far from the median. Finally, the corrected values are included in T_{ij} like a weighted mean via

$$T_{bi} = \frac{\sum_{j=1}^{n} w(u_j) x_j}{\sum_{j=1}^{n} w(u_j)}. \tag{2.4}$$

Here, n is the total sample size of x. In MAS 5.0 method, SB_i^+ cannot be less than or equal to 0. In order to guarantee its positivity, a threshold point γ is taken into account by using the median of the distribution of $(\log \mathrm{PM}_{ij} - \log \mathrm{MM}_{ij})$. Accordingly, if $SB_i^+ > \gamma$, no further adjustment is made to the data, otherwise, the following equation is applied in place of the original SB_i^+:

$$SB_i^+ \quad = \quad \frac{\gamma}{1 + 0.1(\gamma - SB_i^+)},$$

which shows a weighting function decreasing to zero slowly.

There are two main disadvantages of MAS 5.0 gene index. These are:

1. It assumes that MM values merely measure the stray signal and the noise additively.
2. The estimation of the true signal can have bias when $\mathrm{PM}_{ij} \leq \mathrm{MM}_{ij}$.

2.2 MBEI(dChip) Method

MBEI (*Model Based Expression Index*) is one of the well-known gene expression indices. As the examples of some microarray applications via this approach, we can consider the study of Guerri et al. (2012), which detects a patient diagnosed with mantle cell lymphonia into the category of indolent or conventional. Janne et al. (2012) use it in an analysis about the cancer treatment with an anti-Erbis therapeutic agent that is related to an activation of MET gene mutation or MET gene amplification. Additionally, Bonner-Weir et al. (2012) apply this method in a study to investigate the glucose-responsive insulin secreting cells in an enriched population of matures and to modulate the insulin expression, activity and secretion in a subject. Then, Shlien and Malkin (2010) and Harbour (2011) perform dChip in the

risk analysis of a mammal having a cancer and in a study to find out the risk of metastasis, respectively.

For the mathematical details of this method, MBEI suggests a multiplicative model for the observed signal and if the dependent and non-identically distributed probe-measures exist, it can take into account the variability of each probe separately (Li and Wong, 2001).

Moreover, it assumes that there is a linearly increasing relation between the intensity of a probe and the model based expression index, θ_a for a gene in the ath array ($a = 1, 2, \ldots, k$). This rate of increase changes in each probe and within the same probe pair. Furthermore, it is accepted that the PM intensity raises faster than the MM intensity. As a result, the full model is described as follows:

$$\begin{aligned} \mathrm{MM}_{aj} &= v_j + \theta_a \alpha_j + \varepsilon_{aj}^m, \\ \mathrm{PM}_{aj} &= v_j + \theta_a \alpha_j + \theta_a \phi_j + \varepsilon_{aj}^p, \end{aligned}$$

in which v_j reports the baseline response resulting from the non-specific hybridization, α_j stands for the rate of increase in stray signal, ϕ_j denotes the additional rate of increase in the PM intensity (true signal, probe efficiency) for the jth probe pair, and both ε_{aj}^m and ε_{aj}^p are the random error of MM and PM, respectively.

Then, the observed probe intensity for the ath array is computed as:

$$Y_{aj} = \mathrm{PM}_{aj} - \mathrm{MM}_{aj} = \theta_a \phi_j + \varepsilon_{aj},$$

where the error term, ε_{aj}, is distributed normally with mean 0 and variance σ^2.

In estimation of the model parameters, the MBEI performs the least square estimation method. Accordingly, the major disadvantage of this index is its limitations working with for large number of arrays. The reason is that the estimates cannot be found explicitly due to the fixed probe effect. However, from the comparative analysis it has been shown that it gives better results than MAS 5.0 in terms of the accuracy of the estimated signals (Purutçuoğlu, 2007; Lemon et al., 2002).

2.3 RMA Method

This method is another very common approach in microarray analysis and often used in a variety of biological researches. For instance Dash et al. (2012) and Greco et al. (2010) use it in the study of gene expression resources for plants and plant pathogens, and study about SV-40 immortalized human corneal epithelial cells cultured with an air-liquid interface, respectively. Moreover, Obayashi et al. (2007) and Le et al. (2012) apply this approach in the analysis to determine the co-regulated gene groups in arabidopsis and to investigate the change in the gene expressions of soybean leaf tissues at the late developmental stages under drought stress, in order.

In the *RMA* (*Robust Microarray Analysis*) method, the intensities from MM probes are considered to include only the non-specific hybridization intensities, thereby, can be disregarded during the estimation (Irizzary et al., 2003).

On the other hand, the intensities from the PM probes are thought to contain both the background and the true signal, which are composed of the optical noise and the non-specific hybridization.

In this method, the following assumptions are accepted to calculate the true signal and get rid of the background signal:

1. It takes into account the conditional expectation of the true signal.
2. It assumes the exponential true signal s_{aij} and the normal background, b_{aij}. Then, the estimated true signal is modeled as below:

$$s^*_{aij} \quad = \quad E(s_{aij}|s_{aij} + b_{aij}) \equiv E(s_{aij}|\text{PM}),$$

where s_{aij} is the true signal and b_{aij} shows the background signal for the ith gene ($i = 1, 2, ..., n$), the ath array ($a = 1, 2, ..., k$), and the jth probe ($j = 1, 2, ..., m$).

Finally, the gene expression index is described by:

$$\log_2 (s^*_{aij}) \quad = \quad \mu_{ai} + \alpha_{ij} + \varepsilon_{aij}, \tag{2.5}$$

in which μ_{ai} shows the expression level for the ath array on the logarithmic scale, α_{ij} presents the jth probe effect, and ε_{aij} indicates the error term with mean 0 for each gene i. In Equation (2.5), before the estimation of underlying model parameters, the outlier probes are detected and the assumption of $\sum_{j=1}^{m} \alpha_j = 0$ is set. Then the method implements the quantile normalization, which represents a transformation of the arrays by setting the same distribution of the probe intensities for each array (Bolstad et al., 2003). Finally, after this array's normalization, the intensities are transformed to the log-scale.

On the other hand, from the comparison of this method with others, there is a similarity between RMA and MBEI methods in terms of modeling the probe effect. However, for individual probe effect, RMA suggests an additive model on the log-scale, whereas, MBEI proposes a multiplicative one on the original scale.

Also, due to the fixed effects model for the probe effects, RMA is not good at working with large number of arrays, similar to MBEI.

Moreover, RMA has some advantages over MAS 5.0 and MBEI with respect to the standard deviations, in particular, for genes at lower intensities across replicated arrays (i.e. it gives smaller standard deviations), and with respect to the higher consistency in fold-change estimates under different concentrations. Furthermore, from the analysis via *ROC* (*Receiving Operating Characteristic*) curves, it can detect the differentially expressed genes better (Purutçuoğlu, 2007).

2.4 mgMOS Method

The *mgMOS* (*Modified Gamma Model for Oligonucleotide Signal*) method is another alternative model for single channel microarrays. In the literature this method is also widely used in biological researches, such that Noyes et. al. (2011) apply it in a study for detecting cattle identified candidate genes in pathways responding to Trypanosoma congolense infection. Then Cusumano et. al. (2010) and Derrien et. al. (2011) implement it in the analysis of prenatal mouse cochlea and virulence plasmid harbored by uropathogenic escherichia coli (i.e., E-coli) in acute stages of pathogenesis, in order. Furthermore, Derrien et. al. (2011) and Buler et. al. (2011) perform this method in the study about the modulation of the mucosal immune response, tolerance and proliferation in mice colonized as well as in a study regarding the energy sensing factors, respectively.

In this model, the MM probe intensities are treated as only coming from the background signal and the true signal is modeled according to a joint probability density generated via a gamma distribution for PM and MM probes (Milo et al., 2003). Different from RMA, in order to model the correlation between the PM and MM intensities, this model makes use of latent variables, standing for different binding affinity of probes within a specified probe set.

In this method, the PM and MM intensities are assumed to be distributed as gamma with the same inverse scale but with different shape parameters. Thereby, the model can be expressed as follows:

$$\gamma_{ij} = m_{ij} + s_{ij}, \tag{2.6}$$

where N is the number of probes on the chip ($j = 1, \ldots, N$) and n_j is the number of probes in the jth probe set ($i = 1, \ldots, n_j$). In Equation (2.6), γ stands for the observed PM intensity, m shows the observed MM intensity, and s presents the true probe signal.

Due to the advantage of the probabilistic model standing for the relationship between the data and model parameters, by this model, it is possible to find credibility intervals for the expression indices (Purutçuoğlu, 2007).

2.5 GC-RMA Method

The GC-RMA (Robust Microarray Analysis based on GC content) model is an extended version of the RMA approach in the sense that it is the first method, which can come up with the idea of presence of true signal intensities in the MM probes (Wu et al., 2004). In the literature, this method is implemented in a variety of analysis such as the study about rhythmic plant growth (Michael et al., 2008), the retinal gene expression in chicks during imposed myopic defocus (Schippert et. al., 2008), Amacrine cell layer of chicks after myopic and hyperopic defocues (Asbby and Feldkaemper, 2010), non-small cell lung carcinoma cell lines (Dalby et al., 2012), coordinated histone modifications (Ha et al., 2011), age and mortality of humans (Kerber et al., 2009) and multiple myeloma (Meibner et al., 2011).

Hereby, in this approach, a fraction term is added in the MM probes to explain the amount of true intensities. However, in practice it is assumed that it can set to zero as the minimum

intensity from an array. On the other side, the PM values are considered to come from three sources, namely, the optical noise, the non-specific binding, and the true signal. The first two sources are treated as independent functions of the probe affinity, which is the sum of position-base effects.

Thus, the underlying model for any particular probe pair can be shown as follows:

$$\text{PM} = O_{\text{PM}} + N_{\text{PM}} + S, \tag{2.7}$$

$$\text{MM} = O_{\text{MM}} + N_{\text{MM}} + \phi S, \tag{2.8}$$

where O shows the optical noise, N stands for the NSB noise and S is a quantity proportional to RNA expression. The term ϕ presents the fact that MM intensities hasve some true values.

On the other hand, regarding the comparison of accuracy among GC-RMA, RMA and MAS 5.0 methods, GC-RMA and MAS 5.0 give better results than RMA, whereas, RMA performs better with respect to the precision (Purutçuoğlu, 2007).

2.6 BGX Method

Different from GC-RMA, the *BGX* (*Bayesian Expression Index*) method does not assume the value of the fraction in Equation (2.7) as zero. Instead, its value is estimated from the data (Hein et al., 2005). The aim of this approach is to reduce the variances of the estimated expression by the help of information from MM probes via the underlying fraction term.

Hereby, in this model, MM probes are considered to contain some fraction of the true signal and the signal from the cross-hybridization. On the other hand, PM probes contain the signal from cross-hybridization and the true signal. Additionally, both of the probe sets are assumed to be normally distributed with the following parameters:

$$\text{PM}_{ij} \sim N(S_{ij} + H_{ij}, \psi^2) \quad \text{and} \quad \text{MM}_{ij} \sim N(\Phi S_{ij} + H_{ij}, \psi^2), \tag{2.9}$$

where S_{ij} refers to the true signal and H_{ij} presents the non-specific and cross-hybridization for probe j of gene i ($i = 1, \ldots, n$). Moreover, ψ^2 denotes the constant variance of each probe while the fraction term lies within 0 and 1.

In the estimation of the model parameters, the Bayesian methods are applied. Accordingly, on the logarithmic scale, the following hierarchical structure is considered:

$$\log(S_{ij} + 1) \sim TN(\mu_i, \sigma_i^2) \quad \text{and} \quad \log(H_{ij} + 1) \sim TN(\varphi, \eta^2),$$

in which μ_i and σ_i^2 display the gene specific hyperparameters of the true signal for the ith gene and jth probe under the truncated normal distribution. Similarly, φ and η^2 stand for the associated hyperparameters of the non-specific and cross-hybridization signal H_{ij} under the

same distribution. There are also some assumptions for the listed hyperparameters, which are listed as below:

$$\begin{aligned} \mu_i &\sim \text{Uniform}(0, 15) \\ \log(\sigma_i^2) &\sim N(e, f^2) \\ \phi &\sim \text{Beta}(1, 1) \\ \varphi &\sim N(0, 1000) \\ (\psi^2)^{-1} \quad &\text{and} \quad (\eta^2)^{-1} \sim \text{Gamma}(0.001, 0.001), \end{aligned}$$

where e and f^2 show the empirical mean and variance of σ_i^2, respectively, in order to describe the variance of $\log(S_{ij} + 1)$ for each probe set i.

Finally, as the point estimator of BGX, the median of the posterior signal distribution is computed.

On the other hand, as the advantages of this method over others, it provides smaller bias, in particular, at low levels of gene expression and it gives better accuracy than MAS 5.0, MBEI, and RMA. Moreover, BGX and RMA are better than the previous methods for ranking genes regarding the extent of the differential expression. However, as its disadvantage, it is computationally intensive (Purutçuoğlu and Wit, 2006; Purutçuoğlu, 2007). Therefore, it has not yet commonly used in biological researches, apart from some recent studies such as a research about type-one diabetes (Jailwala et al., 2009) and chondrogenic differentiation of human bone marrow-derived mesenchymal stem cells (Herlofsen et al. 2011).

2.7 Multi-mgMOS Method

The *Multi-mgMOS* (*Multiple Chips mgMOS*) method suggests the same model of BGX as shown in Equation (2.9), whereas, it uses the Bayesian estimation with a maximum likelihood approximation in inference of the model parameters to reduce the computational cost of BGX (Hubbell et al., 2002). In the application, this index is used all the biological analyses presented for the mgMOS method.

Hereby, in this model, the binding fraction is estimated from the empirical knowledge gathered from the spike-in genes with concentrations higher than 50 picoMolar (pM). Moreover, for highly expressed spike-in genes, the background and the non-specific hybridization are assumed to be zore.

In the parameter estimation, making use of maximum a posteriori estimate (MAP) under the log-normal prior, the logarithm of the posterior probability is maximized. To find the index for gene i on array a, the median of the expected log true probe signals across all probe pairs are used. The distribution of these true signals depends on some probe-set specific parameters, b_i, d_i, and the posterior distribution of array specific parameter α_{ai}.

Although, Multi-mgMOS is computationally more efficient than BGX, the cost of the computation is still demanding for large datasets since this model makes use of the Bayesian approach, whereas, it gives as sensitive results as BGX finds (Purutçuoğlu, 2007).

2.8 FGX Method

From the Q-Q (quantile-quantile) plots of Affymetrix genes, it is found that a correlation between the values of PM and MM probes exists on the logarithmic scale (Purutçuoğlu and Wit, 2006). The reason for this linear relation can be caused by the existence of some gene-specific target values on the MM values. Indeed as stated in GC-RMA, BGX, and multi-mgMOS methods, both PM and MM values contain this signal, resulting in a significant correlation between PM and MM values.

Such correlation implies that PM and MM values possess a part of a common signal S. In addition, in both PM and MM values, there exists a large non-specific hybridization component μ_H, as an off-set term. On the other hand, the assumption of the log-normality gives the opportunity to deal with the heterogeneity of the variance across the intensity range.

Then, the corresponding model is presented as follows:

$$\log \mathrm{PM}_{ij} \sim N(S_i + \mu_H, \sigma^2)$$

and

$$\log \mathrm{MM}_{ij} \sim N(pS_i + \mu_H, \sigma^2), \tag{2.10}$$

where S_i corresponds to the true expression value for gene i, p stands for the fraction of specific hybridization to the mismatch probe and μ_H denotes the mean of the non-specific signal, containing the non-specific hybridization, background, and the stray signal. Moreover, i and j are the gene and probe indicators, respectively, where $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

On the other side, in Equation (2.10), σ^2 refers to the constant variance that is composed of the nested variance of measurement error and background signal.

In inference of model parameters, because of the fact that the averages of $\log(PM)$ and $\log(MM)$ values are sufficient statistics for the corresponding means and the analysis of the Affymetrix data is done on a probe set, rather than on individual, level, the following averages are considered in calculation.

$$\mathrm{PM}_i = \sum_{j=1}^{m} \frac{log\mathrm{PM}_{ij}}{m} \text{ and } \mathrm{MM}_i = \sum_{j=1}^{m} \frac{log\mathrm{MM}_{ij}}{m},$$

such that

$$\mathrm{PM} \sim N(S_i + \mu_H, \frac{\sigma^2}{m}) \text{ and } \mathrm{MM} \sim N(pS_i + \mu_H, \frac{\sigma^2}{m}). \tag{2.11}$$

However, the expressions in Equation (2.11) are not sufficient statistics for estimating the variance terms. One way to deal with this challenge can be to reformulate the likelihood function in terms of all data after the estimation of the remaining parameters, S_i, p, and μ_H and then to compute the MLE for σ^2 conditional on the estimates $\hat{S}_i$, $\hat{p}$, and $\hat{H}$ by using the invariance property of maximum likelihood estimation (MLE) (Purutçuoğlu, 2007). In the biological literature, this method is not implemented often yet, apart from boron toxicity in a sensitive barley cultivar leaves (Purutçuoğlu et al., 2012), but, it is performed different gene-expression based comparative studies (Kennedy, 2008; Augugliaro and Mineo, 2010; Sarmah and Samarasinghe, 2011; Purutçuoğlu, 2012).

2.9 RGX Method

The RGX (Robust Gene Expression Index) is a recently developed method, which can be considered as the extension of the FGX approach (Purutçuoğlu, 2012) and in application via real life data, it is used in the analysis of the boron toxicity of the barley leaves. In this method, different from FGX, we can handle non-normal log-expressions, in particular, long-tailed symmetric densities.

Accordingly, in RGX, the logarithms of PM and MM are thought to come from long-tailed symmetric families (LTS). But under LTS, since the partial derivatives of log-likelihood functions do not have explicit solutions for estimates of parameters, this method makes use of the modified maximum likelihood estimation (MMLE) technique in inference.

On the other hand, the reason for the selection of LTS, rather than normality, can be observed from the Q-Q plot of PM and MM probe values of the different microarray datasets. In these data, it is seen that there exists deviations from the straight line, especially, in the tails, which can be seen as the indictions of the underlying density. Hereby, the PM and MM in this model are described as below:

$$\begin{aligned} \log \mathrm{PM}_{ij} &\sim \mathrm{LTS}(S_i + \mu_H, \sigma^2) \\ \log \mathrm{MM}_{ij} &\sim \mathrm{LTS}(pS_i + \mu_H, \sigma^2), \end{aligned}$$

whose model parameters are derived from the MMLE via:

$$\begin{aligned} \hat{\mu}_H &= \frac{\sum_{j=1}^{m} \beta_j \mathrm{MM}_i - \hat{p} \sum_{j=1}^{m} \beta_j \mathrm{PM}_i}{(p-1)\sum_{j=1}^{m} \beta_j}, \\ \hat{\sigma} &= \frac{B + \sqrt{B^2 + 4nmC}}{2nm}, \end{aligned}$$

where

$$\begin{aligned} B &= \frac{v}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_j (\mathrm{PM}_{ij} - \mathrm{MM}_{ij}), \\ C &= \frac{v}{k} \left[\sum_{i=1}^{n} \sum_{j=1}^{m} (\mathrm{PM}_{ij} - \hat{S}_i - \hat{\mu}_H)^2 + \sum_{i=1}^{n} \sum_{j=1}^{m} (\mathrm{MM}_{ij} - \hat{p}\hat{S}_i - \hat{\mu}_H)^2 \right] \end{aligned}$$

and

$$\hat{S}_i = \frac{\hat{\sigma}(1+\hat{p})\alpha_j + (\mathrm{PM}_{ij} + \hat{p}\mathrm{MM}_{ij})\beta_j - \hat{\mu}_H(1+\hat{p})\beta_j}{(1+\hat{p})\beta_j}.$$

In the estimation of $\hat{p}$, the index requires a two-stage procedure, where in the first stage, all the estimates are found based the least square estimate (LSE) of p, i.e., under normality assumption. Once $\hat{\mu}_H$, $\hat{S}_i$ and $\hat{\sigma}$ are computed, they are used in the MML estimator of p from the invariance property of the MML method. Finally, all estimates are run three or four times so that all the concommitants fix in the iteration.

From the comparative analysis, it has shown that RGX gives better results than its well known alternatives in the detection of differentially expressed genes. Also, RGX is more computationally efficient than FGX and under the non-normality, it outperforms FGX as well (Purutçuoğlu, 2012; Purutçuoğlu et al., 2012).

In the following part, we present our novel gene expression index, called as *multi-RGX* with the derivation of its estimator. Moreover, we describe other alternative modellings of our index, give their derivations and discuss their challenges.

CHAPTER 3

multi-RGX METHOD

As described in the previous chapters and stated in the Introduction part, the major aim of our study is to suggest a new gene expression index, which is as fast as its strong alternatives such as FGX and RGX, overcome the problems of computational challenge of BGX model whose inference is based on the bayesian approach, and reach high accuracy in the estimated signals without ignoring probe level information in the oligonucleotides like in the modelling of FGX and RGX.

Accordingly, in general, the multi-RGX (multi-Robust Gene Expression Index) can be seen as the extended version of the RGX method. Thereby, in this model, we suppose that the intensities from PM and MM probes on the log-scale have the density from the long-tailed symmetric families, similar to the idea of RGX. On the other hand, different from RGX, it does not follow an iterative process, but it works with the explicit solutions of the desired parameter estimates. In fact, we also consider possible other modelling approaches of multi-RGX by including signal level variance, fraction, and background signal at different levels. Whereas, as described in the following section in this chapter, none of these alternatives can give explicit solutions in estimation and since one of the major goals of our study is to gain from the computational time against the full modelling of the BGX approach based on the iterative algorithms, we do not use these alternative modellings in the application.

In a standard Affymetrix GeneChip, there are m number of probes, and in each probe there exists n number of genes coming from PM and MM probes.

As mentioned previously in FGX and RGX indices, we also try to model the intensities coming from MM probes since it contains not only the false signal, but also a fraction of the true signal. To define the amount of the underlying true signal, a fraction term p is applied. Hereby, the intensities coming from PM probes is distributed as the long-tailed symmetric with mean containing a constant term and a term, which stands for the true signal changing from gene to gene. On the other hand, its variance is constant over genes and probes. Different from PM, MM has the mean containing the same constant term and the part of the true signal.

In the estimation of the model parameters, the modified maximum likelihood estimation (MMLE) method is applied whose procedure is described as follows.

In finding the MLE, there are some situations where we cannot find the solution(s) of the likelihood equation explicitly (Tiku and Akkaya, 2004). Under such challenges, rather than applying iterative methods, Tiku has proposed an alternative approach, called the *modified*

maximum likelihood estimation (MMLE), which can solve the nonlinearity problems in the partial derivatives functions of model parameters in the underlying likelihood functions.

Basically, the MML method considers replacement of the intractable terms in the partial derivative of the log-likelihood equations by their linear approximations via the first order Taylor series expansion. The procedure for this approach can be described under a location-scale family of the symmetric distribution, which has the form below (Tiku and Akkaya, 2004; Akkaya and Tiku, 2007):

$$f(x) \quad = \quad \frac{1}{\sigma\sqrt{k}\beta(\frac{1}{2}, v-\frac{1}{2})}\left[1+\frac{(X-\mu)^2}{k\sigma^2}\right]^{-v}, \; -\infty < x < \infty,$$

where $k = 2v - 3$ and the shape parameter v is greater than or equals to 2 ($v \geq 2$) to guarantee the existence of the expectation μ, where $E(X) = \mu$ and $V(X) = \sigma^2$. Finally, $\beta(\frac{1}{2}, v-\frac{1}{2}) = \frac{\Gamma(\frac{1}{2})\Gamma(v-\frac{1}{2})}{\Gamma(v)}$, in which $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ and $\Gamma(v) = (v-1)!$.

In the estimation, considering that $X_1, \ldots, X_n$ are a random sample of size n and v is known, the corresponding likelihood function can be written as:

$$L(X_1, X_2, \ldots, X_n|\mu, \sigma^2) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{k}\beta(\frac{1}{2}, v-\frac{1}{2})}\left[1+\frac{(X_i-\mu)^2}{k\sigma^2}\right]^{-v}, \tag{3.1}$$

$$= \quad \left[\frac{1}{\sigma\sqrt{k}\beta(\frac{1}{2}, v-\frac{1}{2})}\right]^{n} \prod_{i=1}^{n}\left[1+\frac{(X_i-\mu)^2}{k\sigma^2}\right]^{-v}. \tag{3.2}$$

Since in Equation (3.2) the term $\frac{1}{\sigma\sqrt{k}\beta(\frac{1}{2}, v-\frac{1}{2})}$ is constant in terms of μ and σ, the function can be simplified via:

$$L(X_1, X_2, \ldots, X_n|\mu, \sigma^2) \quad \propto \quad \frac{1}{\sigma^n}\prod_{i=1}^{n}\left[1+\frac{(X_i-\mu)^2}{k\sigma^2}\right]^{-v}$$

and the log-likelihood can be written proportionally by:

$$\ln L \quad \propto \quad n\ln\sigma - v\sum_{i=1}^{n}\left[1+\frac{(X_i-\mu)^2}{k\sigma^2}\right]. \tag{3.3}$$

Then, by substituting X_i as $z_i = \frac{X_i-\mu}{\sigma}$, Equation (3.3) can be described as:

$$\ln L \quad \propto \quad -n\ln\sigma - v\sum_{i=1}^{n}\left[1+\frac{1}{k}z_i^2\right].$$

Finally, the maximum likelihood (ML) estimators of μ and σ are derived from the following partial derivatives:

$$\frac{\partial \ln L}{\partial \mu} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \frac{z_i}{1 + \frac{1}{k} z_i^2} = \frac{2p}{k\sigma} \sum_{i=1}^{n} g(z_i) \tag{3.4}$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^{n} \frac{z_i^2}{1 + \frac{1}{k} z_i^2} = -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^{n} z_i g(z_i). \tag{3.5}$$

Since by setting $g(z_i) = \frac{z_i}{1+\frac{z_i^2}{k}}$, Equations (3.4) and (3.5) do not have explicit solutions, the desired estimates can be found by MML, rather than MLE, method. In this approach, we order the data from smallest to largest magnitude and write Equations (3.4) and (3.5) via:

$$\frac{\partial \ln L}{\partial \mu} = \frac{2p}{k\sigma} \sum_{i=1}^{n} g(z_{(i)})$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^{n} z_{(i)} g(z_{(i)}),$$

respectively, where $z_{(i)} = \frac{X_{(i)} - \mu}{\sigma}$ and $X_{(i)}$ presents the ith order statistic for $i = 1, \ldots, n$.

As we need to linearize $\frac{\partial ln L}{\partial \mu}$ and $\frac{\partial \ln L}{\partial \sigma}$ in order to get explicit solutions, we apply the Taylor series of the nonlinear function $g(z_{(i)})$ as follows:

$$\begin{aligned}
g(z_{(i)}) &\cong g(t_{(i)}) + (z_{(i)} - t_{(i)}) \left[g'(z) \right] \\
&= \frac{t_{(i)}}{1 + \frac{t_{(i)}^2}{k}} + (z_{(i)} - t_{(i)}) \frac{1 + \frac{z_{(i)}^2}{k} - z_{(i)} \frac{1}{k} 2 z_{(i)}}{1 + \frac{1}{k} z_{(i)}^2} \\
&= \frac{t_{(i)}}{1 + \frac{t_{(i)}^2}{k}} + (z_{(i)} - t_{(i)}) \frac{1 + \frac{t_{(i)}^2}{k} - z_{(i)} \frac{1}{k} 2 t_{(i)}}{1 + \frac{1}{k} t_{(i)}^2} \\
&= \frac{t_{(i)}}{1 + \frac{t_{(i)}^2}{k}} + z_{(i)} \frac{1 - \frac{t_{(i)}^2}{k}}{(1 + \frac{t_{(i)}^2}{k})^2} - t_{(i)} \frac{1 - \frac{t_{(i)}^2}{k}}{(1 + \frac{t_{(i)}^2}{k})^2} \\
&\cong \alpha_i + \beta_i z_{(i)},
\end{aligned}$$

in which $z_{(i)}$ denotes the ordered z_i with respect to gene i ($i = 1, \ldots, n$) as the concommitant while

$$\alpha_i = \frac{2t_{(i)}^3}{k} \text{ and } \beta_i = \frac{1 - \frac{t_{(i)}^2}{k}}{(1 + \frac{t_{(i)}^2}{k})^2}$$

under $\sum_{i=1}^{n} \alpha_i = 0$ due to the symmetry. Here, $t_{(i)}$ stands for the ordered student-t values for each gene i.

By this way, we can obtain a linear equation for $g(z_{(i)})$ in place of its nonlinear expression. Then, we get the MML estimators from the following modified likelihood equations $\partial \ln L^*/\partial\mu$ and $\partial \ln L^*/\partial\sigma$:

$$\frac{\partial \ln L}{\partial \mu} \approx \frac{\partial ln L^*}{\partial \mu} = \frac{2p}{k\sigma} \sum_{i=1}^{n} \alpha_i + \beta_i z_{(i)} \tag{3.6}$$

$$\frac{\partial \ln L}{\partial \sigma} \approx \frac{\partial \ln L^*}{\partial \sigma} = -\frac{n}{\sigma} + \frac{2p}{k\sigma} \sum_{i=1}^{n} z_{(i)}(\alpha_i + \beta_i z_{(i)}). \tag{3.7}$$

Thus, by solving Equation (3.6) and (3.7) simultaneously and equating them to zero, we derive:

$$\begin{aligned}
\sum_{i=1}^{n} \alpha_i + \beta_i z_{(i)} &= 0 \\
\sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \beta_i z_{(i)} &= 0 \\
\sum_{i=1}^{n} \beta_i \frac{(X_{(i)} - \mu)}{\sigma} &= 0 \\
\sum_{i=1}^{n} \beta_i (X_{(i)} - \mu) &= 0 \\
\sum_{i=1}^{n} \beta_i X_{(i)} - \sum_{i=1}^{n} \beta_i \mu &= 0.
\end{aligned}$$

Hence, the estimate for μ is obtained as:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \beta_i X_{(i)}}{\sum_{i=1}^{n} \beta_i}.$$

Similarly,

$$
\begin{aligned}
-\frac{n}{\sigma}+\frac{2p}{k\sigma}\sum_{i=1}^{n}(z_{(i)}\alpha_i+\beta_i z_{(i)}^2) &= 0\\
-\frac{1}{\sigma^3}\left[n\sigma^2-\frac{2p\sigma^2}{k}\sum_{i=1}^{n}z_{(i)}(\alpha_i+\beta_i z_{(i)})\right] &= 0\\
n\sigma^2-\frac{2p\sigma^2}{k}\sum_{i=1}^{n}\frac{(X_{(i)}-\hat{\mu})}{\sigma}(\alpha_i+\beta_i\frac{(X_{(i)}-\hat{\mu})}{\sigma}) &= 0\\
n\sigma^2-\frac{2p\sigma^2}{k}\sum_{i=1}^{n}\left[\frac{(X_{(i)}-\hat{\mu})\alpha_i}{\sigma}+\beta_i\frac{(X_{(i)}-\hat{\mu})^2}{\sigma^2}\right] &= 0\\
n\sigma^2-\frac{2p\sigma}{k}\sum_{i=1}^{n}((X_{(i)}-\hat{\mu})\alpha_i+\beta_i\frac{(X_{(i)}-\hat{\mu})^2}{\sigma}) &= 0\\
n\sigma^2-\frac{2p\sigma}{k}\sum_{i=1}^{n}(X_{(i)}-\hat{\mu})\alpha_i+\frac{2p}{k}\sum_{i=1}^{n}\beta_i(X_{(i)}-\hat{\mu})^2 &= 0\\
n\sigma^2-\sigma(\frac{2p}{k}\sum_{i=1}^{n}X_{(i)}\alpha_i-\frac{2p}{k}\hat{\mu}\sum_{i=1}^{n}\alpha_i+\frac{2p}{k}\sum_{i=1}^{n}\beta_i(X_{(i)}-\hat{\mu})^2) &= 0\\
n\sigma^2-\sigma\frac{2p}{k}\sum_{i=1}^{n}X_{(i)}\alpha_i+\frac{2p}{k}\sum_{i=1}^{n}\beta_i(X_{(i)}-\hat{\mu})^2 &= 0.
\end{aligned}
$$

Thereby, the estimate for σ is found as:

$$\hat{\sigma}=\frac{B+\sqrt{B^2+4nC}}{2\sqrt{n(n-1)}},$$

where

$$B=\frac{2p}{k}\sum_{i=1}^{n}X_{(i)}\alpha_i \text{ and } C=\frac{2p}{k}\sum_{i=1}^{n}\beta_i(X_{(i)}-\hat{\mu})^2.$$

Here n is replaced by $\sqrt{n(n-1)}$ to reduce the bias.

On the other hand, the procedure for $g(z)$ yields the same result obtained from ML estimation if $g(z)$ is linear. Moreover, it has been shown that the limits of both Equations (3.4) and (3.6) as well as Equations (3.5) and (3.7) are equivalent such that

$$
\begin{aligned}
\lim_{n\to\infty}\frac{\frac{\partial \ln L}{\partial\mu}}{n} &\equiv \lim_{n\to\infty}\frac{\partial \ln L^*}{\partial\mu}\equiv 0\\
\lim_{n\to\infty}\frac{\frac{\partial \ln L}{\partial\sigma}}{n} &\equiv \lim_{n\to\infty}\frac{\partial lnL^*}{\partial\sigma}\equiv 0.
\end{aligned}
$$

These results imply that the MML estimates are asymptotically equal to ML estimates. Accordingly, $\hat{\mu}$ and $\hat{\sigma}$ derived from MML method are asymptotically unbiased and efficient (Tiku et al., 1986; Tiku and Akkaya, 2004).

In the following parts, we initially present the summary of derivations of multi-RGX and then describe other possible modelling approaches with their derivations because of the fact that none of them can produce explicit forms for their estimators. The detailed derivations for each model are given ,n Appendices.

3.1 Derivation of the multi-RGX Estimators

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we consider the following distributional assumption on the log-scale:

$$\mathrm{PM}_{ij} = a_{ij} \sim \mathrm{LTS}(S_i + \mu_H, \sigma^2)$$

and

$$\mathrm{MM}_{ij} = b_{ij} \sim \mathrm{LTS}(pS_i + \mu_H, \sigma^2),$$

where LTS denotes the long-tailed symmetric density.

Thereby, the corresponding likelihood is found via:

$$\begin{aligned} L(S_i, \mu_H, p \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a_{ij}-S_i-\mu_H)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(b_{ij}-pS_i-\mu_H)^2}{2\sigma^2}}, \end{aligned}$$

which is proportional to

$$L \propto (\frac{1}{\sigma}) \prod_{i=1}^{n}\prod_{j=1}^{m} (1 + \frac{z_{a_{ij}}^2}{k})^{-v} (\frac{1}{\sigma}) \prod_{i=1}^{n}\prod_{j=1}^{m} (1 + \frac{z_{b_{ij}}^2}{k})^{-v},$$

where $v \geq 2$, $k = 2v - 3$.

In order to calculate the MLE of the model parameters, by making use of common nonlinear functions stated in previous section, the first derivatives of the function $\ln L$ are taken with respect to each parameter as below:

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma}}{\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2}} + \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\frac{(b_{ij}-pS_i-\mu_H)}{\sigma}}{\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}} \\
\frac{\partial \ln L}{\partial p} &= \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\frac{S_i(b_{ij}-pS_i-\mu_H)}{\sigma}}{\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}} \\
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma}}{\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2}} + \frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{-p(b_{ij}-pS_i-\mu_H)}{\sigma}}{\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}} \\
\frac{\partial \ln L}{\partial \sigma} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma}+\frac{(a_{ij}-S_i-\mu_H)}{\sigma^3(1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2})}\frac{2v}{k}\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma}+\frac{(b_{ij}-pS_i-\mu_H)}{\sigma^3(1+\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2})}\frac{2v}{k}\right],
\end{aligned}
$$

By first order Taylor expansions followings can be driven:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}}.$$

In these expressions, $z_{a_{i(j)}}$ and $z_{b_{i(j)}}$ show the ordered probes (in increasing magnitude) for each gene i in PM and MM standardized intensities, respectively. Accordingly, $g(z_{a_{i(j)}})$ and $g(z_{b_{i(j)}})$ are their associated linearized function via the Taylor series. Thereby, the partial derivatives of MLE are as follows:

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} g(z_{b_{i(j)}}) \\
\frac{\partial \ln L}{\partial p} &= \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} S_i g(z_{b_{i(j)}}) \\
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k\sigma}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2vp}{k\sigma}\sum_{j=1}^{m} g(z_{b_{i(j)}}) \\
\frac{\partial \ln L}{\partial \sigma} &= -\frac{2nm}{\sigma} + \frac{2v}{k\sigma}\sum_{j=1}^{m} g(z_{a_{i(j)}}) z_{a_{i(j)}} + \frac{2v}{k\sigma}\sum_{j=1}^{m} g(z_{b_{i(j)}}) z_{b_{i(j)}}.
\end{aligned}
$$

By making some substitutions $\hat{\mu}_H$ can be written in a way shown below:

$$\hat{\mu}_H = \frac{p\sum_{j=1}^{m}\beta_j \bar{a}_{.j} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j}}{(p-1)\sum_{j=1}^{m}\beta_j}.$$

Similarly, the form of $\hat{S}_i$ can be derived as below:

$$\begin{aligned}\hat{S}_i &= \frac{(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})p^2}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)} \\ &+ \frac{(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j} + \sum_{j=1}^{m}\beta_j \bar{b}_{.j})p}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)} \\ &+ \frac{\sum_{j=1}^{m}\beta_j \bar{b}_{.j} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j},\end{aligned}$$

for $\bar{b}_{.j} = \sum_{i=1}^{n} \frac{a_{i(j)}}{n}$.

On the other hand, the estimate of the common fraction term p can be found as below:

$$\begin{aligned}\hat{p}_1 &= \hat{p}_2 = 1 \\ \hat{p}_3 &= \frac{(SS_b - SS_a) - \sqrt{(SS_a - SS_b)^2 + 4SS_{ab}^2}}{2SS_{ab}} \\ \hat{p}_4 &= \frac{(SS_b - SS_a) + \sqrt{(SS_a - SS_b)^2 + 4SS_{ab}^2}}{2SS_{ab}},\end{aligned}$$

where

$$\begin{aligned}SS_a &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j}\right]^2 \\ SS_b &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j}\right]^2 \\ SS_{ab} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j})\right].\end{aligned}$$

Finally, the MML estimate for σ can be presented by:

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4nmC}}{nm},$$

for S_i $(i = 1, \ldots, n)$. By adjusting the degree of freedom as $2nm - (n+2)$, the estimate of σ can be indicated as follows:

$$\hat{\sigma} = \frac{B + \sqrt{B^2 + 4nmC}}{nm - n - 2},$$

in which

$$
\begin{aligned}
B &= \frac{v}{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j(a_{i(j)}-b_{i(j)})\right],\\
C &= \frac{v}{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)}-\hat{S}_i-\hat{\mu}_H)^2+\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(b_{i(j)}-\hat{p}\hat{S}_i-\hat{\mu}_H)^2\right].
\end{aligned}
$$

The full derivation for all estimators can be found in Appendix A. The MMLE estimators are also efficient and the covariances-variances of model can be found via the Inverse of the Fisher Information matrix (Tiku et al., 1986; Tiku and Akkaya, 2004). In the following section, we present the derivation of all covariance and variance estimators for our model parameters, which are developed from the observed Fisher information matrix.

3.2 Observed Fisher Information Matrix and Estimators for Variances and Covariances

The Fisher Information Matrix is generated by making use of the second partial derivatives of the loglikelihood function with respect to each parameter as below:

$$
\begin{aligned}
I_{11} &= -\frac{\partial^2 l}{\partial {\mu^2}_H}\\
I_{22} &= -\frac{\partial^2 l}{\partial p^2}\\
I_{ii} &= -\frac{\partial^2 l}{\partial S_i^2}\\
I_{12} &= I_{21} = -\frac{\partial^2 l}{\partial \mu_H \partial p}\\
I_{1i} &= I_{i1} = -\frac{\partial^2 l}{\partial S_i \partial \mu_H}\\
I_{2i} &= I_{i2} = -\frac{\partial^2 l}{\partial S_i \partial p}\\
I_{ik} &= I_{ki} = -\frac{\partial^2 l}{\partial S_i \partial S_k} = 0,
\end{aligned}
$$

where $i, k = 1, \ldots, n$ present the gene and $j = 1, \ldots, m$ shows the probe indicator, respectively. Moreover, l stands for $\ln L$, the loglikelihood function.

Then, the Fisher information matrix is derived as below:

$$I = \begin{bmatrix} \frac{\partial^2 l}{\partial \mu^2{}_H} & \frac{\partial^2 l}{\partial \mu_H \partial p} & \frac{\partial^2 l}{\partial \mu_H \partial S_1} & \cdots & \frac{\partial^2 l}{\partial \mu_H \partial S_n} \\ \frac{\partial^2 l}{\partial p \partial \mu_H} & \frac{\partial^2 l}{\partial p^2} & \frac{\partial^2 l}{\partial p \partial S_1} & \cdots & \frac{\partial^2 l}{\partial p \partial S_n} \\ \frac{\partial^2 l}{\partial S_1 \partial \mu_H} & \frac{\partial^2 l}{\partial S_1 \partial p} & \frac{\partial^2 l}{\partial S_1^2} & ... & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 l}{\partial S_n \partial \mu_H} & \frac{\partial^2 l}{\partial S_n \partial p} & 0 & ... & \frac{\partial^2 l}{\partial S_n^2}. \end{bmatrix}$$

In order to find the variance-covariance matrix we take the advantage of the above information matrix and derive the following variance and covariance terms:

$$\begin{aligned} V(\hat{\mu}_H) &= \frac{1}{C_0}\left[\frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} S_i^2 \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right. \\ &- \left.\sum_{i=1}^{n} \frac{-\frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \end{aligned}$$

$$\begin{aligned} V(\hat{p}) &= \frac{1}{C_0}\left[\frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right]\right. \\ &- \left.\sum_{i=!}^{n} \frac{(\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2})^2}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \end{aligned}$$

$$
\begin{aligned}
Cov(\hat{\mu}_H, \hat{p}) &= \left[\frac{2v}{k\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m} S_i \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2} \right. \\
&+ \sum_{i=1}^{n} \frac{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2}} \\
&\times \left(-\frac{2v}{k\sigma} \sum_{j=1}^{m} \frac{\frac{(b_{i(j)} - pS_i - \mu_H)}{\sigma}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} \right. \\
&+ \left. \left. \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} S_i \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2} \right) \right] \times \frac{1}{C_0},
\end{aligned}
$$

where

$$
\begin{aligned}
C_0 &= \left(-\frac{\partial^2 l}{\partial \mu_H^2} - \sum_{i=1}^{n} \frac{(\partial^2 l / \partial S_i \partial \mu_H)^2}{-\partial^2 l / \partial S_i^2} \right) \times \left(-\frac{\partial^2 l}{\partial p^2} - \sum_{i=1}^{n} \frac{(\partial^2 l / \partial S_i \partial p)^2}{-\partial^2 l / \partial S_i^2} \right) \\
&- \left(-\frac{\partial^2 l}{\partial p \partial \mu_H} - \sum_{i=1}^{n} \frac{(-\partial^2 l / \partial S_i \partial \mu_H)(-\partial^2 l / \partial S_i \partial p)}{-\partial^2 l / \partial S_i^2} \right).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
Cov(\hat{S}_i, \hat{\mu}_H) &= \left[\frac{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times -V(\hat{\mu}_H) \\
&- \left[\frac{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(a_{i(j)} - S_i - \mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1 - \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}}{\left(1 + \frac{(b_{i(j)} - pS_i - \mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{\mu}_H, \hat{p})
\end{aligned}
$$

and

$$
\begin{aligned}
Cov(\hat{S}_i, \hat{p}) &= \left[\frac{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times -Cov(\hat{\mu}_H, \hat{p}) \\
&- \left[\frac{-\frac{2v}{k\sigma} \sum_{j=1}^{m} \frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times V(\hat{p})
\end{aligned}
$$

Also,

$$
\begin{aligned}
V(\hat{\hat{S}}_i) &= 1 - \left[\frac{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{\hat{S}}_i, \hat{\mu}_H) \\
&- \left[\frac{-\frac{2v}{k\sigma} \sum_{j=1}^{m} \frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2} \sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2} \sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{\hat{S}}_i, \hat{p}).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
Cov(\hat{S}_i, \hat{S}_k) &= -\left[\frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \\
&\times Cov(\hat{S}_k, \hat{\mu}_H) \\
&- \left[\frac{-\frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp}{k\sigma^2}\sum_{j=1}^{m}S_i\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \\
&\times Cov(\hat{S}_k, \hat{p}).
\end{aligned}
$$

The derivation of each variance and covariance structure is presented in Section A.1 under Appendix A. In the next chapters, we evaluate the performance of our model by comparing the outputs of benchmark and simulated dataset whose estimated signals are already found by other alternative approaches. For the calculation, we generate our own R codes as a new function. The codes are also presented in Appendix F. But previously we derive plausible modelling approaches of multi-RGX and show that all these models are based on iterative techniques since none of them has close-form estimators.

3.3 Alternative Models of multi-RGX Index

Apart from the derivation given in the previous sections, we also derive the estimators of possible alternative of multi-RGX in this part. In those models, we add the gene specific variance (Alternative Model 1), the gene specific variance and fraction (Alternative Model 2), the gene specific variance and background signal (Alternative Model 3), and finally, the gene specific variance, fraction as well as background signal (Alternative Model 4), respectively. Whereas, as seen in the corresponding sub-sections, none of these models produces explicit expressions for model parameters, resulting in iterative approach in calculation. Therefore, we choose the current model given in the previous section for multi-RGX as its estimators are close and explicit forms and also, these estimators are asymptotically equivalent to MLE results even though our selected model is based on simpler assumptions than its alternatives. Summary of the derivations is given here and the details can be found in Appendices.

3.3.1 Alternative Model 1

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we consider the following distributional assumption (long-tailed symmetric LTS distribution) on the log-scale:

$$\begin{aligned}\mathrm{PM}_{ij} = a_{ij} &\sim \mathrm{LTS}(S_i + \mu_H, \sigma_i^2),\\ \mathrm{MM}_{ij} = b_{ij} &\sim \mathrm{LTS}(pS_i + \mu_H, \sigma_i^2).\end{aligned}$$

Here S_i is the true signal for gene i, μ_H refers to the constant background intensity, and σ_i presents the gene specific standard deviation. Finally, p indicates the fraction of the true signal in MM.

Accordingly, the associated likelihood function is found via:

$$\begin{aligned}L(S_i, \mu_H, p \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij})\\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_H)^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-pS_i-\mu_H)^2}{2\sigma_i^2}},\end{aligned}$$

which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m}\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m}\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v},$$

where the shape parameter $v \geq 2, k = 2v-3, a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$.

In order to get the MLE of model parameters, we take the first derivatives of lnL with respect to each parameter as follows:

$$\begin{aligned}\frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} + \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}}\\ \frac{\partial \ln L}{\partial p} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i(b_{ij}-pS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}}\\ \frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} - \frac{2vp}{k}\sum_{j=1}^{m}\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}}\\ \frac{\partial \ln L}{\partial \sigma_i} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2})}\frac{2v}{k}\right]\\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i^3(1+\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2})}\frac{2v}{k}\right].\end{aligned}$$

Then, by making use os common nonlinear functions and by the first order Taylor expansions we get the following equations:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}}$$

and

$$g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}}.$$

when $a_{i(j)}$ and $b_{i(j)}$ represent the ordered PM and MM with respect to the probes j, i.e., the concomitant. Hereby, by using the partial derivatives of MMLE with respect to each parameter and by taking $\sum_{j=1}^{m} \alpha_j = 0$ due to the symmetry, we obtain the following equations:

$$\begin{aligned}
\hat{\mu}_H &= \frac{p\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{(ij)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{(ij)}}{(p^2+1)\sum_{j=1}^{m}\beta_j}, \\
\hat{S}_i &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p^2+1)\sum_{j=1}^{m}\beta_j} - \frac{\frac{p(p+1)}{(p-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{i(j)} - \frac{(p+1)}{(p-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}}{(p^2+1)\sum_{j=1}^{m}\beta_j}.
\end{aligned}$$

Hereby,

$$\hat{p} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i}{\sigma_i^2}b_{i(j)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\mu_H\frac{\beta_j S_i}{\sigma_i^2}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i^2}{\sigma_i^2}}.$$

Also,

$$\begin{aligned}
\hat{\sigma_i}^2 &= \frac{km}{v}\left[\frac{p(p-1)}{(p^2+1)}\sum_{j=1}^{m}\beta_j a_{i(j)} - \frac{(p-1)}{(p^2+1)}\sum_{j=1}^{m}\beta_j b_{i(j)}\right. \\
&- \left.\frac{p(p-1)}{(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{i(j)}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\right] + \frac{km}{v}\left[\frac{(p-1)}{(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\right].
\end{aligned}$$

We present the complete derivation of the alternative model 1 in Appendix B.

3.3.2 Alternative Model 2

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, the following distributional assumption under the long-tailed symmetric(LTS) density is considered on the log-scale:

$$\mathrm{PM}_{ij} = a_{ij} \sim \mathrm{LTS}(S_i + \mu_H, \sigma_i^2)$$

and

$$\mathrm{MM}_{ij} = b_{ij} \sim \mathrm{LTS}(p_i S_i + \mu_H, \sigma_i^2),$$

where S_i and μ_H are the gene specific true signal and background intensity, respectively, and σ_i denotes the standard deviation for gene i as used in previous alternative models. On the other hand, p_i presents the fraction of the true signal in MM for each gene i.

Thereby, the corresponding likelihood is found via:

$$\begin{aligned} L(S_i, \mu_H, p_i \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_H)^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-p_i S_i-\mu_H)^2}{2\sigma_i^2}}, \end{aligned}$$

which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m} (\frac{1}{\sigma_i})(1 + \frac{z_{a_{ij}}^2}{k})^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m} (\frac{1}{\sigma_i})(1 + \frac{z_{b_{ij}}^2}{k})^{-v},$$

where the shape parameter $v \geq 2$, $k = 2v - 3$, degree of freedom $d = 2v - 1$, and finally a and b refer to nm-dimensional vectors $a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$, in order.

To obtain the MLE of model parameters, we take the first derivatives of $\ln L$ with respect to each parameter as below:

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(a_{ij} - S_i - \mu_H)}{\sigma_i^2 \frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} + \frac{2v}{k} \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(b_{ij} - p_i S_i - \mu_H)}{\sigma_i^2 \frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}} \\ \frac{\partial \ln L}{\partial p_i} &= \frac{2v}{k} \sum_{j=1}^{m} \frac{S_i(b_{ij} - p_i S_i - \mu_H)}{\sigma_i^2 \frac{(b_{ij}-p_i S_i-\mu_H)^2}{k\sigma_i^2}} \\ \frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k} \sum_{j=1}^{m} \frac{(a_{ij} - S_i - \mu_H)}{\sigma_i^2 \frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} + \frac{2v}{k} \sum_{j=1}^{m} \frac{-p_i(b_{ij} - p_i S_i - \mu_H)}{\sigma_i^2 \frac{(b_{ij}-p_i S_i-\mu_H)^2}{k\sigma_i^2}} \\ \frac{\partial \ln L}{\partial \sigma_i} &= -\sum_{i=1}^{n}\sum_{j=1}^{m} \left[-\frac{1}{\sigma_i} + \frac{(a_{ij} - S_i - \mu_H)}{\sigma_i^3 (1 + \frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2})} \frac{2v}{k} \right] \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{m} \left[-\frac{1}{\sigma_i} + \frac{(b_{ij} - p_i S_i - \mu_H)}{\sigma_i^3 (1 + \frac{(b_{ij}-p_i S_i-\mu_H)^2}{k\sigma_i^2})} \frac{2v}{k} \right]. \end{aligned}$$

Then, we approximate the common nonlinear functions $g(z) = \frac{z}{1+\frac{z^2}{k}}$ for PM and MM as, and we get the following first order Taylor expansions:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

for the standardized and ordered PM and MM intensities (in increasing magnitude), in order, with respect to the probes in each gene i. Moreover, here $t_{(j)}$ refers to the ordered associate student-t quantile for each probe j ($j = 1, \ldots, m$).

Then, the partial derivatives of MLE can be shown as follows:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \\
\frac{\partial \ln L}{\partial p} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \\
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{j=1}^{m}\frac{p_i}{\sigma_i}g(z_{b_{i(j)}}) \\
\frac{\partial \ln L}{\partial \sigma_i} &= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{b_{i(j)}}).
\end{aligned}$$

Accordingly, the form of $\hat{\mu}_H$ can be derived as below:

$$\begin{aligned}
\hat{\mu}_H &= \frac{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\sum_{i=1}^{n}\beta_j} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j a_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}} \\
&\quad - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)p_i}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j b_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}.
\end{aligned}$$

Likewise,

$$
\begin{aligned}
\hat{S}_i &= \frac{1}{(p_i^2+1)}\frac{\sum_{j=1}^m \beta_j a_{i(j)}}{\sum_{j=1}^m \beta_j} + \frac{p_i}{(p_i^2+1)}\frac{\sum_{j=1}^m \beta_j b_{i(j)}}{\sum_{j=1}^m \beta_j} \\
&- \frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\left(2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)\sum_{j=1}^m \beta_j} \\
&- \frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{p_i+1}{(p_i^2+1)}\frac{\beta_j}{\sigma_i} a_{i(j)}}{\left(2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)} \\
&- \frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{p_i(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i} b_{i(j)}}{\left(2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)}.
\end{aligned}
$$

Thereby,

$$
\hat{p} = \frac{\sum_{j=1}^m b_{i(j)} - \mu_H \sum_{j=1}^m \beta_j}{S_i \sum_{j=1}^m \beta_j}.
$$

Hence, the form of the variance term can be derived via:

$$
\begin{aligned}
\hat{\sigma_i}^2 &= \frac{km}{v}\sum_{j=1}^m \beta_j (a_{i(j)}+b_{i(j)}) \\
&- \frac{(p_i+1)v}{km}\left[\frac{1}{(p_i^2+1)}\frac{\sum_{j=1}^m \beta_j a_{i(j)}}{\sum_{j=1}^m \beta_j} + \frac{p_i}{(p_i^2+1)}\frac{\sum_{j=1}^m \beta_j b_{i(j)}}{\sum_{j=1}^m \beta_j}\right] \\
&+ \frac{(p_i+1)v}{km}\left[\frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{(2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})\sum_{j=1}^m \beta_j}\right] \\
&+ \frac{(p_i+1)v}{km}\left[\frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{p_i+1}{(p_i^2+1)}\frac{\beta_j}{\sigma_i} a_{i(j)}}{(2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})}\right] \\
&+ \frac{(p_i+1)v}{km}\left[\frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{p_i(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i} b_{i(j)}}{(2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})}\right] \\
&- 2\sum_{j=1}^m \beta_j \left[\frac{\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\sum_{i=1}^n \beta_j} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^n \beta_j a_{i(j)}}{2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}\right] \\
&+ 2\sum_{j=1}^m \beta_j \left[\frac{\sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)p_i}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^n \beta_j b_{i(j)}}{2\sum_{i=1}^n \frac{1}{\sigma_i} - \sum_{i=1}^n \sum_{j=1}^m \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}\right] = 0.
\end{aligned}
$$

From this final expression, it is seen that similar to the first alternative model, Model 2 also has none explicit expression for the model estimators. The full derivation of this model is given in Appendix C.

3.3.3 Alternative Model 3

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we consider the following distributional assumption under the long tailed symmetric (LTS) on the log-scale:

$$\begin{aligned} \text{PM}_{ij} &= a_{ij} \sim \text{LTS}(S_i + \mu_{Hi}, \sigma_i^2), && (3.8) \\ \text{MM}_{ij} &= b_{ij} \sim \text{LTS}(pS_i + \mu_{Hi}, \sigma_i^2), && (3.9) \end{aligned}$$

in which S_i, p and μ_H describe the true signal in gene i, constant fraction of true signal in MM, and gene-specific background intensities, respectively, as described beforehand. Finally, σ_i^2 is the gene-specific variance component.

Hereby, the associated likelihood is found via:

$$\begin{aligned} L(S_i, \mu_{Hi}, p \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_{Hi})^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{2\sigma_i^2}}, \end{aligned}$$

for $a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$, which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m} \left(\frac{1}{\sigma_i}\right)\left(1 + \frac{z_{a_{ij}}^2}{k}\right)^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m} \left(\frac{1}{\sigma_i}\right)\left(1 + \frac{z_{b_{ij}}^2}{k}\right)^{-v}$$

and under the shape parameter $v \geq 2$ and $k = 2v - 3$.

In order to get the MLE of model parameters, we take the first derivatives of $\ln L$ with respect

to each parameter as follows:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_{Hi}} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{j=1}^{m}\frac{(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}}\\
\frac{\partial \ln L}{\partial p} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}}\\
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{j=1}^{m}\frac{-p(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}}\\
\frac{\partial \ln L}{\partial \sigma_i} &= \sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right]\\
&+ \sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right].
\end{aligned}$$

Then, by approximating the common nonlinear functions $g(z) = \frac{z}{1+\frac{z^2}{k}}$ for PM and MM, the first order Taylor expansions can be written via:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

where

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2},$$

under the probe based ordered values of z_{ij} for each gene i. Here $t_{(i)}$ indicates the quantile of the student-t density for the jth probe, as used other alternative models. We can express the partial derivatives of MMLE as below:

$$\frac{\partial \ln L}{\partial \mu_{Hi}} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}})+\frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \tag{3.10}$$

$$\frac{\partial \ln L}{\partial p} = \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \tag{3.11}$$

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}})+\frac{2vp}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \tag{3.12}$$

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{2m}{\sigma_i}+\frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{a_{i(j)}})+\frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{b_{i(j)}}). \tag{3.13}$$

Finally, $\hat{\mu}_H i$ is found as:

$$\hat{\mu}_{Hi} = \frac{p\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{j=1}^{m}\beta_j b_{(ij)}}{(p^2+1)\sum_{j=1}^{m}\beta_j}.$$

Similarly, $\hat{S}_i$ can be written as:

$$\hat{S}_i = \frac{\sum_{j=1}^{m} \beta_j b_{i(j)} - \sum_{j=1}^{m} \beta_j a_{i(j)}}{(p-1)\sum_{j=1}^{m} \beta_j}.$$

Thereby,

$$\hat{p} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\beta_j S_i}{\sigma_i^2} b_{i(j)} - \sum_{i=1}^{n}\sum_{j=1}^{m} \mu_{Hi} \frac{\beta_j S_i}{\sigma_i^2}}{\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\beta_j S_i^2}{\sigma_i^2}}.$$

Finally, we can obtain MLE of σ for each gene i as follows:

$$\hat{\sigma_i}^2 = \frac{\sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p+1) S_i \sum_{j=1}^{m} \beta_j - 2\mu_{Hi} \sum_{j=1}^{m} \beta_j}{\frac{km}{v}}.$$

Then, we get the estimate of $\hat{\sigma_i}$ as below:

$$\frac{km\hat{\sigma_i}^2}{v} = 0.$$

Hereby,

$$\hat{\sigma_i} = 0,$$

which implies an infeasible estimate for σ_i. We describe the derivation of estimators for each model parameter in Appendix D.

3.3.4 Alternative Model 4

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we assume the following relation under long-tailed symmetric (LTS) distribution on the log-scale:

$$\text{PM}_{ij} = a_{ij} \sim \text{LTS}(S_i + \mu_{Hi}, \sigma_i^2), \tag{3.14}$$

$$\text{MM}_{ij} = b_{ij} \sim \text{LTS}(p_i S_i + \mu_{Hi}, \sigma_i^2), \tag{3.15}$$

for gene specific true signal S_i, background intensity μ_{Hi}, and variance σ_i^2.

Thereby, the associated likelihood can be written as follows:

$$L(S_i, \mu_{Hi}, p_i \mid a, b) = \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij})$$

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij} - S_i - \mu_{Hi})^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij} - p_i S_i - \mu_{Hi})^2}{2\sigma_i^2}}$$

under $a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$, which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m}(\frac{1}{\sigma_i})(1+\frac{z_{a_{ij}}^2}{k})^{-v}\prod_{i=1}^{n}\prod_{j=1}^{m}(\frac{1}{\sigma_i})(1+\frac{z_{b_{ij}}^2}{k})^{-v},$$

for the shape parameter $v \geq 2$ and $k = 2v - 3$.

In order to get the MLE of model parameters, we take the first derivatives of $\ln L$ with respect to each parameter as below:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_{Hi}} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}} + \frac{2v}{k}\sum_{j=1}^{m}\frac{(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial \ln L}{\partial p_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{S_i(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}} + \frac{2v}{k}\sum_{j=1}^{m}\frac{-p_i(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial \ln L}{\partial \sigma_i} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right].
\end{aligned}$$

Moreover, by approximating the common nonlinear functions $g(z) = \frac{z}{1+\frac{z^2}{k}}$ for PM and MM, we can find the first order Taylor expansions v,a

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

as well as

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2},$$

By using the partial derivatives of MLE as follows, we can find the forms of estimates as below:

$$\hat{\mu}_{Hi} = \frac{p_i\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{j=1}^{m}\beta_j b_{(ij)}}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}.$$

Similarly,

$$\hat{S}_i = \frac{\sum_{j=1}^{m} \beta_j b_{i(j)} - \sum_{j=1}^{m} \beta_j a_{i(j)}}{(p_i - 1)\sum_{j=1}^{m} \beta_j}.$$

On the other side, we can get the estimate of the common fraction term p_i as below:

$$\hat{p} = \frac{\sum_{j=1}^{m} b_{i(j)} - \mu_{Hi} \sum_{j=1}^{m} \beta_j}{S_i \sum_{j=1}^{m} \beta_j}.$$

Above equation gives us no solution, which implies an infeasible estimate for p_i.

Then, the variance term is derived as below:

$$\begin{aligned}
\frac{km\hat{\sigma}_i^2}{v} &= \sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p_i + 1)\left[\frac{-\sum_{j=1}^{m} \beta_j a_{i(j)} + \sum_{j=1}^{m} \beta_j b_{i(j)}}{(p_i - 1)\sum_{j=1}^{m} \beta_j}\right] \\
&- 2\sum_{j=1}^{m} \beta_j \left[\frac{p_i \sum_{j=1}^{m} \beta_j (a_{(ij)} - \sum_{j=1}^{m} \beta_j (b_{(ij)})}{(p_i^2 + 1)\sum_{j=1}^{m} \beta_j}\right] \\
&= \sum_{j=1}^{m} \beta_j a_{i(j)} + \sum_{j=1}^{m} \beta_j b_{i(j)} - \frac{p_i + 1}{p_i - 1}\sum_{j=1}^{m} \beta_j b_{i(j)} + \frac{p_i + 1}{p_i - 1}\sum_{j=1}^{m} \beta_j a_{i(j)} \\
&- \frac{2p_i}{p_i - 1}\sum_{j=1}^{m} \beta_j a_{i(j)} + \frac{p_i + 1}{p_i - 1}\sum_{j=1}^{m} \beta_j a_{i(j)} - \frac{p_i}{p_i - 1}\sum_{j=1}^{m} \beta_j a_{i(j)} \\
&= \left[1 + \frac{p_i + 1}{p_i - 1} - \frac{2p_i}{p_i - 1}\right]\sum_{j=1}^{m} \beta_j a_{i(j)} + \left[1 - \frac{p_i + 1}{p_i - 1} + \frac{2}{p_i - 1}\right]\sum_{j=1}^{m} \beta_j b_{i(j)},
\end{aligned}$$

since

$$1 + \frac{p_i + 1}{p_i - 1} - \frac{2p_i}{p_i - 1} = 0 \text{ and } 1 - \frac{p_i + 1}{p_i - 1} + \frac{2}{p_i - 1} = 0.$$

Hereby, we obtain 0 as the estimate of the variance term σ_i as below:

$$\begin{aligned}
\frac{km\hat{\sigma}_i^2}{v} &= 0 \\
\hat{\sigma}_i &= 0,
\end{aligned}$$

which implies infeasible estimator for the standard deviation. The complete derivation for all estimators can be found in Appendix E.

CHAPTER 4

Application

4.1 Application via Real Datasets

4.1.1 Description of Real Datasets

In the analysis, we use two benchmark datasets, namely, dataset 1, which is the benchmark Affymetrix spike-in data, and dataset 2, which is gathered from the GeneLogic spike-in data. The first dataset, i.e., dataset 1, is available at `http://affycomp.biostat.jbsph/edu/`. This dataset consists of 11 spike-in genes: AFFX-DapX-3, AFFX-DapX-5, AFFX-BioC-5, AFFX-DapX-M, AFFX-CreX-3, AFFX-BioC-3, AFFX-BioB-5, AFFX-CreX-5, AFFX-BioB-M, AFFX-BioB-3, and AFFX-BioDn-3 coming from four bacterial ancestor genes. There are 59 arrays with 10864 probe pairs in this set.

For the analysis, the 16 spike-in probe pairs, which are used previously in other comparative analyses are taken. Hereby, the selected spike-in probe pairs are numbered as 3777, 684, 1597, 38734, 39058, 36311, 36889, 1024, 36202, 36085, 40322, 407, 1091, 1708, 33818, and 546. Moreover, in this dataset, the gene expression values of the individual cRNA fragments, which are hybridized to U95A GeneChip arrays at the same concentration are analyzed under 14 different concentration levels. The chosen concentration levels are 0.0, 0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0, 128.0, 256.0, 512.0, and 1024.0 pM (picoMolar). Accordingly, every gene is described by 16 probes in each array. This dataset contains 176 (16 genes and 11 probes per genes) number of observations for each array.

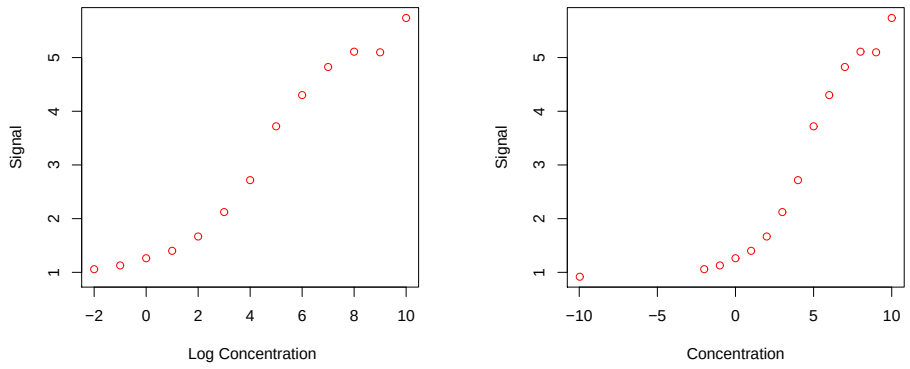
On the other hand, the second dataset, i.e., dataset 2, contains 11 genes, namely, BioB-5, BioB-M, BioB-3, BioC-5, BioC3, BoiDn-3, DapX-5, DapX-M, DapX-3, CreX-5, and CreX-5 (with affix AFFX) in which each gene has 20 probes. Thereby in each array there are 220 (11 genes and 20 probes per gene) observations. Furthermore, 14 arrays, called as 92453, 92454, 92456, 92458, 92460, 92464, 92466 and 92491-92496 (with 9 suffix hgu95a11) are used in the analysis. In this set of measurements, all spike-in genes, except CreX-3, on each array are spiked-in at different concentration levels, which are 0.0, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 5.0, 12.5, 25.0, 50.0, 75.0, 100.0, and 150.0 pM and are composed of 20 probes.

4.1.2 Analysis of Affymetrix Dataset

In the analysis, we use the R programme language (version 2.0.1) because of the compatibility of some packages as the data can be readable under this version of R. Then, we write our

newly developed codes as presented in Appendix. Accordingly, in the calculation, we implement *affy* and *hgu95acdf* R libraries to read the `.cel` files in all arrays. These two libraries are specifically designed to read the data in Affymetrix and GeneLogic `hgu` files.

In the estimation of the model parameters for each dataset, we check the possible shape parameters for the long-tailed symmetric (LTS) distribution from 2 to 40 with 0.05 jump size. The value, which maximizes the log-likelihood under the least square estimators is chosen as the selected shape parameter of LTS for the taken array. From this calculation, we find that most of the array can be defined with shape parameter 40 or close to 40, which indicates close to normal distribution in practice for this Affymetrix dataset.



(a) (b)

Figure 4.1: (a) Average estimated signal versus concentrations on the nominal logarithmic scale ($\log_2$) for the dataset 1 via multi-RGX method and (b) the same plot by excluding 0 concentration and corresponding estimated signal.

In Figure 4.1, we plot the average estimated signals versus nominal, i.e., $\log_2$, log-concentrations values. As presented in these plots, while the level of concentrations increases, a nonlinear pattern is easier to be seen in the structure of the estimated signals. In fact the same pattern is also observed from the plots of other methods. Figure 4.2 shows the associated graph via all other methods. For the estimated of RGX, as we find exactly the same structure of FGX as represented in Figure 4.2(b), we do not draw it in a separate graph. On the other hand, in Figure 4.1(b), we plot the same figure of 4.1(a) by replacing the zero concentration with 0.001 concentration since $\log_2(0) = -\infty$ in the analysis.

Moreover, the estimated signals via multi-RGX versus nominal log-concentrations are plotted for each array and the graph is reported in Figure 4.3.

Among all types of concentrations, we separate their levels into three groups, namely low, medium and high concentrations. Hereby, the ones under 0.5, 1.0, 2.0, 4.0, 8.0 and 16.0 pM are defined as low concentrations, the signals under 32.0, 64.0, 128.0, and 256.0 pM are described as medium concentrations, and finally, the concentrations 512.0 and 1024.0 pM are expressed as high concentrations. The reason of such grouping is that the intensities under low to high concentrations indicate distinct behaviours in the sense that the signals under low and high concentrations possess more noisy signals, thereby, are more used to show nonlinear

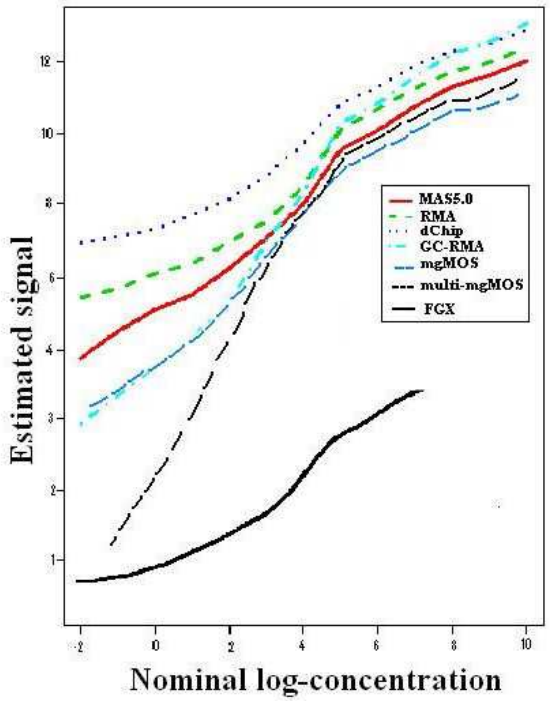


Figure 4.2: Average estimated signals of versus nominal log-concentrations of the dataset 1 (excluding 0.0 pM concentration) using MAS 5.0, dCHIP, RMA, GC-RMA, mgMOS, multi-mgMOS, and FGX method.

and random pattern. On the other hand, under medium levels, the signals can be less affected by the erroneous intensities and can be described via more linear pattern (Hein et al., 2005; Purutçuoğlu and Wit, 2006; Purutçuoğlu, 2012). We show the average estimated signal under each concentration type versus their associated nominal log-concentrations in Figure 4.4.

Then, in order to evaluate the performance of estimated line under each level of concentrations, we apply the simple linear regression method. In the calculation, due to the fact that some concentrations are measured more than once in each array, the mean values of those concentrations are computed and assigned as their common concentrations. Furthermore, the estimated signals under zero pM concentration are omitted. The associated coefficients of determination R^2 for all levels and separate groups are listed in Table 4.4. In this table the computed R^2 is called as *the signal detect* R^2 and its associated slope term is named as the *signal detect slope*. These values are one of the major comparison criteria to asses different gene indices (Cope, 2003; Purutçuoğlu, 2012; Purutçuoğlu, 2007). On the other hand, for the comparison based on the simulated dataset, we assess mainly the relative efficiency between the most competitive approach of multi-RGX, namely, FGX and RGX, since their other scores in real datasets are very close to each other. More details about this comparison can be found in Chapter 5.

On the other hand, if the normalization was implemented within an ANOVA model as described in Chapter 2, different model selection criteria could be applicable. For instance, in order to test the significance of main and interaction effects in Equation2.2, Ülgen (2010) performs W-test, as an alternative of F-test under non-normal density of errors. Moreover,

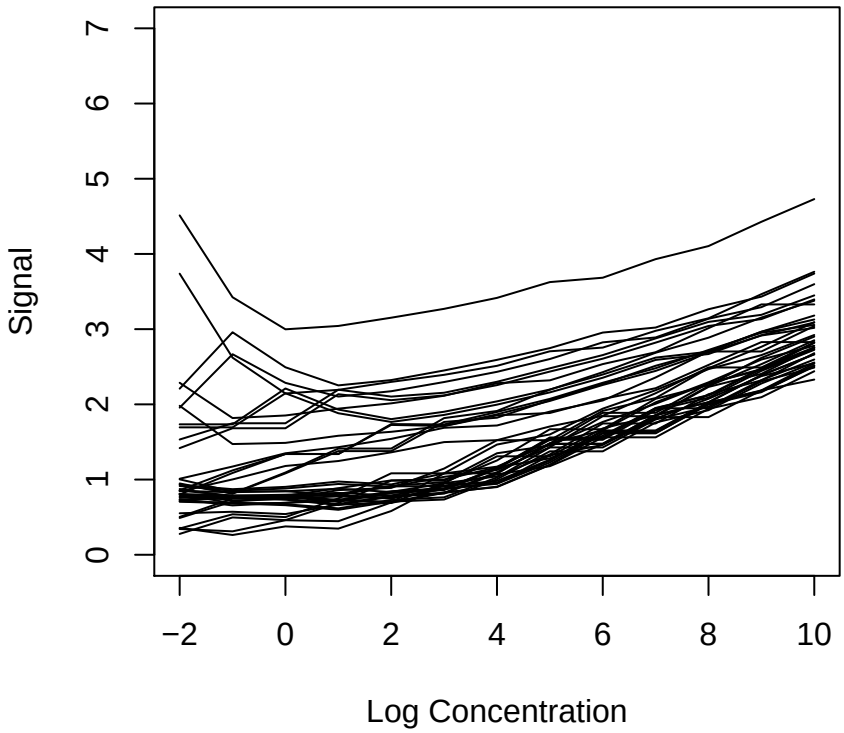
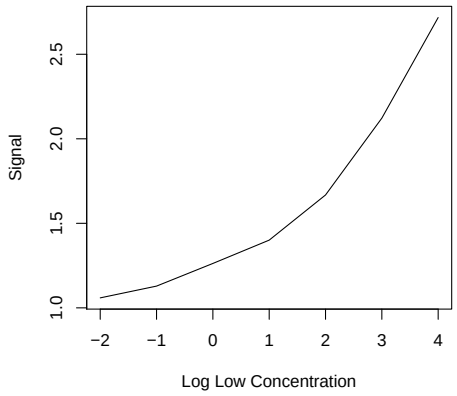
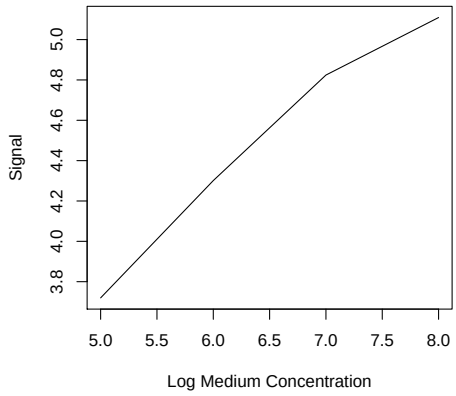


Figure 4.3: Signal versus Log Concentrations for dataset 1

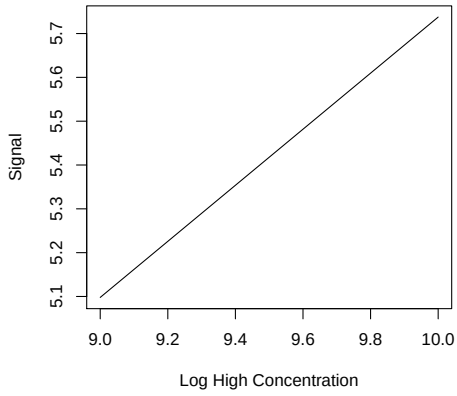
to detect the cause of rejection between the treatment means across treatment/variety in the same equation (Equation 2.2), a modified version of pairwise multiple comparison t-test under non-normality (Dunnett, 1982) is applied. This test uses the MML estimators of the location and scale parameters for the model in Equation 2.2) in place of the original expressions in the Dunnett pairwise multiple comparison test (Ülgen, 2010). Finally, in order to evaluate the robustness of estimators between the W-test versus F-test and modified t-test versus Dunnett t-test, the estimators found from MMLE and AMMLE methods are compared under different scenarios from misspecification and Dixon's mixture to contamination models based on their type 1 errors and powers. Although such types of model selection criteria can be applicable within an ANOVA based microarray analysis, they cannot be used if the researchers follow the first approach, which implies the preprocessing steps in advance of the actual analysis. Because none of the method presented under background normalization can prepare the whole data for the analysis without any further calculation. Accordingly, in order to detect such significance of the genes under different treatments, the analysis of differentially expressed genes is performed after the full steps of normalization and the quality control checks of the calculations (Wit and McClure, 2004; Steen, 2002; Stekel, 2003). Additionally, as not all the background models suggest the same terms in their models, the value estimated one model cannot be directly comparable with a similar estimator in different model. For example, the estimated signals under RMA are already normalized via *between and within arrays normal-*



(a)



(b)



(c)

Figure 4.4: Average estimated signals under (a) low, (b) medium, and (c) high levels of concentration on nominal log-scale of multi-RGX for the dataset 1.

ization (via the quantile normalization method (Bolstad et al., 2003)). Whereas, none of the other gene expression indices uses the quantile-normalized data in their analysis. On the other hand, not all of the suggested methods in this study applies a distribution assumption in their analysis. For instance, MAS 5.0 and RMA models do not apply any distribution assumption and their calculations are based on robust estimators. However, other methods like BGX, FGX and RGX methods use some distributional assumptions in their computations.

Table 4.1: Estimated signal detect R^2's and their associated slope values for all levels of concentrations together and separately for the dataset 1.

Case	Signal detect R^2	Signal detect slope
All Concentrations	0.74	1.68
Low Concentrations	0.28	1.08
Medium Concentrations	0.70	1.55
High Concentrations	0.06	0.10

Then, we compute the R^2's values and slope terms for each gene separately and later calculate the mean of both values individually. The estimated statistics are presented in Table 4.4. From both Table 4.3 and Table 4.2, it is seen that the performance of multi-RGX under gene based evaluation is good and indicates a linear structure on average. Whereas, the estimated signals under each level of concentration is still not convincing to accept that the relation between signals versus concentration even under medium group is linear. In Table 4.3, we present the associated values of other methods for comparison (Purutçuoğlu, 2012; Cope, 2003).

Table 4.2: Average estimated signal R^2's and their associated slope value for each gene for the dataset 1.

Case	R^2	Slope
All Concentrations	0.92	0.45

In the final assessment, we calculate the array based results for the data and the give the mean of these 59 arrays in Table 4.4. From the result, it is seen that since each array indicates its own behaviour and the outlier in the data can be more effective in the evaluation, none of the

Table 4.3: Signal detect R^2, signal detect slope and average R^2, respectively, for the dataset 1 with their perfection values (Purutçuoğlu, 2012)

Method	Signal detect R^2	Signal detect slope	R^2
Perfection value	1.00	1.00	1.00
MAS 5.0	0.86	0.71	0.89
RMA	0.80	0.63	0.99
MBEI (dChip)	0.85	0.53	0.99
GC-RMA	0.84	0.97	0.99
mgMOS	0.82	0.76	0.96
multi-mgMOS	0.80	1.03	0.96
FGX	0.94	0.43	0.90
RGX	0.96	0.44	0.92

separate level of concentrations shows linear relationship. Whereas, under all concentrations simultaneously, the array indicates relatively more linear relation between estimated signal and concentration on the nominal log-scale on the average.

Table 4.4: Average estimated signal R^2's and their associated slope value for each array for the dataset 1.

Case	R^2	Slope
All Concentrations	0.79	1.04
Low Concentrations	0.46	0.26
Medium Concentrations	0.55	0.26
High Concentrations	0.02	-0.03

On the other hand, in order to compare the computational demand for alternative approaches whose real time scores are stored, we check the real computation time of BGX, multi-mgMOS, and FGX methods and add the results of multi-RGX. The findings are presented in Table 4.5. From the table it is seen that multi-RGX is significantly fast than BGX and multi-mgMOS. But with respect to FGX's time, it is relatively slower. The reason is that we need to compute the concommitant of probe sets, which converge to the true order in two or three iterations and this computation causes extra time that is found in tabulated values. Moreover, the calculation of FGX is based on the mean probe level, on the contrary, the computation of multi-RGX depends on both probe and gene specific values, which complicate the calculation.

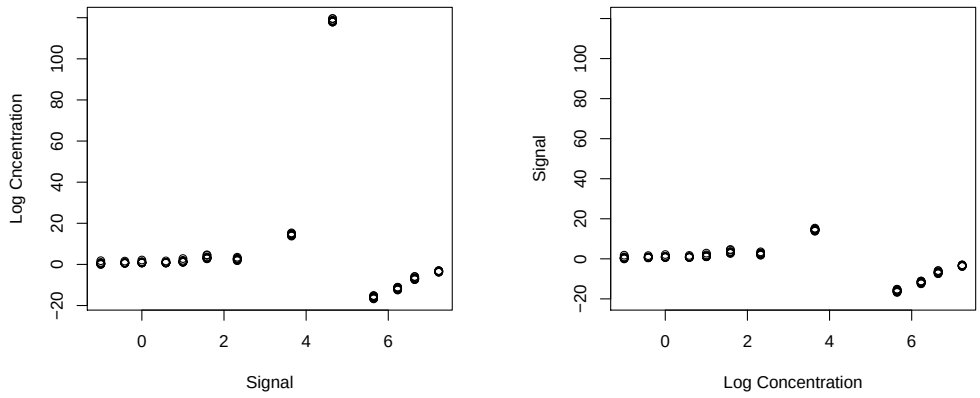
Table 4.5: Real computational time of BGX, multi-mgMOS, FGX, and multi-RGX, respectively, for the dataset 1.

Model	Programme language	Computational time
BGX	C++	32.5 hr
multi-mgMOS	R and C	50 min
FGX	R	4 sec
multi-RGX	R	34 sec

4.1.3 Analysis of GeneLogic Dataset

In the analysis of the GeneLogic dataset, namely, dataset 2, we estimate the model parameters from 13 arrays, in which each array is observed under single concentration as described previously. In the analysis, we initially draw the plots of estimated signal versus concentrations on the nominal log-scale as seen in Figure 4.5(a). From the plot we observe that the signals under 25 pM behave as outliers. In Figure 4.5(b) we draw the same plot by excluding 25 pM and associated intensities.

Then, as implemented in previous analysis, we separate, in particular, the low concentrations, namely, 0.5 and 0.75 pM since the noisy signals become more effective on the observed and estimated signals. In Figure 4.6, we plot the remaining estimated signals versus concentrations on the logarithmic scale. On the other hand, the results found via BGX, MAS 5.0, MBEI, and RMA taken Figure 8 in (Hein et al., 2005) are displayed in Figure 4.7. From their compar-



(a) (b)

Figure 4.5: (a) Average estimated signal versus all concentrations and (b) excluding the 25 pM concentration and associated signals on the nominal logarithmic scale for the dataset2 via multi-RGX method.

isons, it is observed that all plots (excluding the low concentrations) indicate similar pattern apart from the scale of the estimates. The reason is that the multi-RGX, as FGX and RGX, can measure the background intensities when no any gene on the array. Therefore, it can infer the true noisy signal when the arrays are empty. Whereas, other methods combine this noisy source of signal to their estimates, resulting in shifting the estimates with the amount μ_H on average. We get similar findings from the comparative analysis of the Affymetrix dataset (Figure 4.1 and Figure 4.2)

Furthermore, from all plots, it is seen that still the changes in estimated signals cannot be explained linearly via the change in concentrations. But different from previous analyses, here the nonlinearly is strong, hereby, any detection of R^2 is not meaningful. Therefore, the major comparison criterion for this dataset is to evaluate the computational time. In Table 4.6, we find that FGX is the fast method among its strong alternative and both RGX and multi-RGX use the same computational demand. But the former has loss of information due to the ignorance of probe specific estimators. As a result, we can conclude that multi-RGX can be accepted one of the promising methods in background normalization regarding its computational time and accuracy of its estimates with explicit and close forms of all estimators.

Table 4.6: Real computational time of BGX, multi-mgMOS, FGX, RGX, and multi-RGX, respectively, for the dataset 2.

Model	Programme language	Computational time
BGX	C++	70 min
multi-mgMOS	R and C	3 min
FGX	R	1 sec
RGX	R	6 sec
multi-RGX	R	6 sec

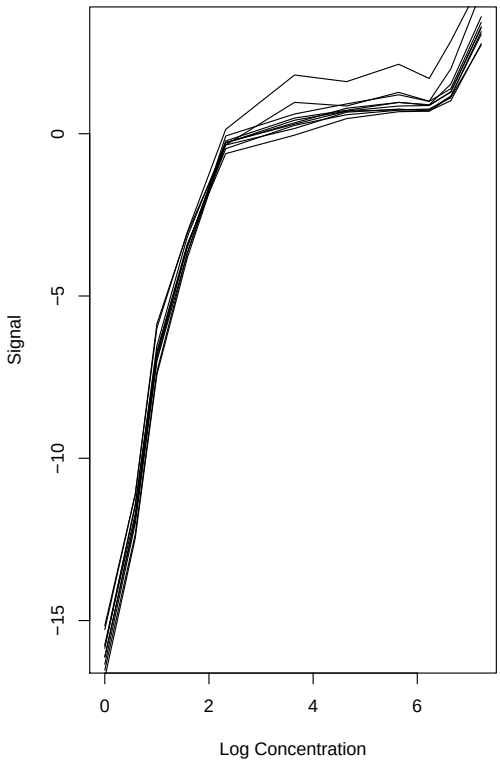


Figure 4.6: Average estimated signal versus concentrations on the nominal logarithmic scale for the dataset 2 via multi-RGX method.

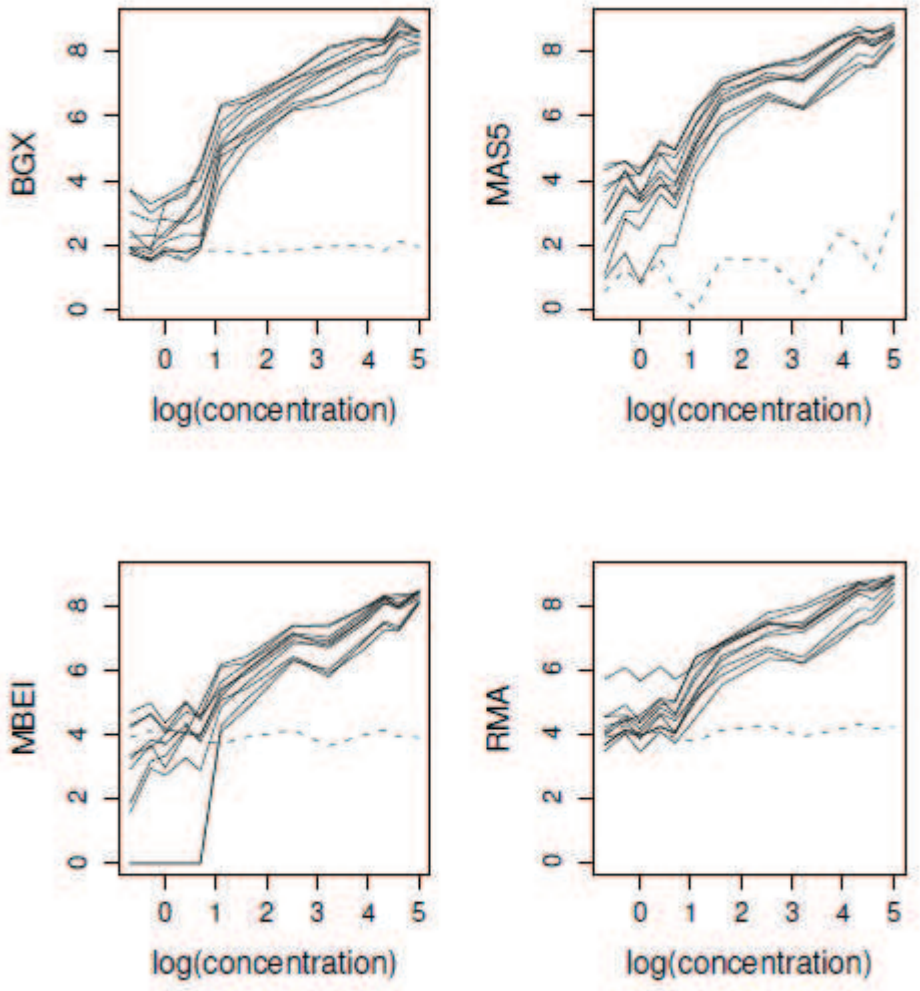


Figure 4.7: Average estimated signal versus concentrations on the nominal logarithmic scale for the dataset 2 via multi-RGX method.

4.2 Application via Simulated Dataset

In order to asses the quality of our novel gene expression index when the data are far from normal or have outliers, we generate, initially, two datasets from 10,000 Monte Carlo runs. Each Monte Carlo run is conducted for 10 genes under 20 probe pairs, resulting in 200 observations for each simulated array. In the first data, we simulate the PM and MM values under normal distribution and in the second data, we generate the data under mean shifted normal density. Then we mix these two sets in order to get two location-mixture datasets with different ratio of outliers. Hereby, the first set is presented as below:

$$0.5N(S_i + \mu_H, \sigma^2) + 0.5N(S_i + \mu_H + \delta\sigma, \sigma^2) \tag{4.1}$$

and the second location mixture has the following structure.

$$0.9N(S_i + \mu_H, \sigma^2) + 0.1N(S_i + \mu_H + \delta\sigma, \sigma^2) \tag{4.2}$$

where $N(.,.)$ indicates the univariate normal with the given parameters. On the other hand, the second data have the following mixture ratio considering that the fist data possess a large number of outliers, whereas, the second one own relatively moderate number of extreme observations. Moreover, in all simulations, we assume that every gene is measured under specific concentration. Thereby, the data are simulated under S_i setting to $S_i = 2, \ldots, 11$ for $i = 1, \ldots, 10$, respectively, on the original scale for each presumed concentration level. Here i stands for the gene indicator. Furthermore, in the simulation μ_H, p, σ are equated as 1, 0.7 and 1, in order. Finally, the shift of location, δ is set to $\delta = 10$ for both datasets.

In the assessment of the estimated results, we select three accuracy criteria, which can be implemented under nonlinear models as well. These three criteria are average error (AE) (Purutçuoğlu and Wit, 2008), mean absolute error (MAE) and root mean square error (RMSE) (Kartal-Koç et al., 2012) whose expressions are presented as below:

$$\begin{aligned} \text{AE} &= \frac{|\hat{\theta}_i - \theta|}{|\theta|}, \\ \text{MAE} &= \frac{1}{n}\sum_{i=1}^{n} |\theta_i - \hat{\theta}_i|, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2}, \end{aligned}$$

where $\hat{\theta}_i$ denotes the estimated model parameters and θ_i is the true value of this model parameter for the gene indicator i ($i = 1, \ldots, n$). Moreover, $|.|$ describes the absolute value of the given term.

In Table 4.7 and Tables 4.8, we list the estimated values, i.e., S_i, μ_H, p and σ, of FGX, RGX and multi-RGX for each gene and model parameters, in order, for the expression in 4.1 and 4.2, respectively, by computing the associated absolute errors.

Table 4.7: Estimated model parameters of the first simulated location-mixture dataset (Expression in 4.1) via FGX, RGX and multi-RGX with their absolute errors (AE) and true values.

Parameter	True value	FGX	AE	RGX	AE	multi-RGX	AE
p	0.7	0.88	0.25	0.87	0.24	0.88	0.25
μ_H	1	-23.91	24.91	-23.91	24.91	-20.99	21.99
σ	1	0.44	0.56	0.27	0.73	0.22	0.79
S_1	2	26.51	12.25	26.50	12.25	23.59	10.79
S_2	3	27.93	8.31	27.93	8.31	23.87	6.96
S_3	4	26.79	5.70	26.78	5.70	24.09	5.02
S_4	5	27.01	4.40	27.00	4.40	24.27	3.85
S_5	6	27.19	3.53	27.19	3.53	24.43	3.07
S_6	7	27.35	2.91	27.35	2.91	24.57	2.51
S_7	8	27.49	2.44	27.48	2.44	24.70	2.09
S_8	9	27.62	2.07	27.61	2.07	24.81	1.76
S_9	10	27.73	1.77	27.72	1.77	24.92	1.49
S_{10}	11	27.84	1.53	27.83	1.53	25.01	1.27

Table 4.8: Estimated model parameters of the second simulated location-mixture dataset (Expression in 4.2) via FGX, RGX and multi-RGX with their absolute errors (AE) and true values.

Parameter	True value	FGX	AE	RGX	AE	multi-RGX	AE
p	0.7	0.95	0.36	0.93	0.33	0.95	0.36
μ_H	1	-25.63	26.63	-25.63	26.63	-22.18	23.18
σ	1	0.61	0.39	0.44	0.56	0.32	0.68
S_1	2	27.33	12.67	27.34	12.67	23.90	10.95
S_2	3	29.29	8.77	29.30	8.77	24.32	7.11
S_3	4	27.76	5.94	27.77	5.94	24.64	5.16
S_4	5	28.08	4.62	28.08	4.62	24.89	3.98
S_5	6	28.34	3.72	28.34	3.72	25.11	3.18
S_6	7	28.55	3.08	28.55	3.08	25.29	2.61
S_7	8	28.73	2.59	28.74	2.59	25.45	2.18
S_8	9	28.90	2.21	28.90	2.21	25.60	1.84
S_9	10	29.04	1.90	29.05	1.91	25.73	1.57
S_{10}	11	29.17	1.65	29.18	1.65	25.85	1.35

Table 4.9: Mean absolute error (MAE) and root mean square error (RMSE) of FGX, RGX and multi-RGX in the calculation of the estimated signals in the two simulated location-mixture datasets.

		MAE	RMSE
First mixture	FGX	20.85	65.92
	RGX	20.84	65.90
	multi-RGX	17.93	56.69
Second mixture	FGX	22.02	69.63
	RGX	22.03	69.65
	multi-RGX	18.58	58.75

From Table 4.7 and 4.8, we observe that there is a large amount of bias in the estimates. There are two reasons of such bias. The first one is that while generating the simulated values for 10 genes, we assume that they are simulated on the original scale with small value of intensities. Then we take $\log_2$ of all the assumed measurements in order to calculate the estimated signals and other model parameters. By this way, indeed, we evaluate the performance of all three modes under the very low and slightly high concentrations because of the fact that the estimates of all indices are problematic, in particular, under the low concentrated values and perform better when the concentrations are high or moderately high. Hereby, we consider that the method which can work well under such challenging range, might work considerably well under non-problematic range of concentrations and intensities. Accordingly, the second reason of high bias is that as observed in Figure 4.2, FGX, RGX and multi-RGX can measure the background intensity when the concentration is zero. In other words, they can calculate the amount of the noisy signals on the array when there is no any effect on the gene. Hence, we can consider the difference between the true and estimated values as the estimated erroreous signals from the surface of the array when the intensities of the genes are really very low as our example.

On the other hand, in Table 4.9, we present the mean absolute error (MAE) and root mean square error (RMSE) for the estimated signals via three selected models and in Table 4.10, we list the real and CPU time of each index in the calculation of the estimated model parameters for these two mixture data.

From all the tabulated values under each model selection criteria, it is seen that the estimates of multi-RGX are more accurate that the ones computed via FGX and RGX indices by implying that our suggested model can improve the accuracy of estimates in highly and moderately fluctuated datasets. On the other hand, the computational demand of multi-RGX is more than other two alternatives in real time. However, this difference is indeed can be tolerable if we compare the results of CPU time. This output shows that with an effective programming the real time of multi-RGX can be even improved since the calculation of the estimates is almost performed under the same computer time in the end.

On the other side, in order to evaluate the performance of each alternative index in a large dataset, we extent the calculation via the simulated data by 10000 and 20000 genes. In both calculations, similar to the first two datasets, we consider that each gene has 20 probes and

Table 4.10: Real and central processing unit (CPU) time (in seconds) of FGX, RGX and multi-RGX in the calculation of the estimated model parameters in the two simulated location-mixture datasets.

		Real time	CPU time
First mixture	FGX	46.44	0.55
	RGX	700.23	0.68
	multi-RGX	2578.98	1.00
Second mixture	FGX	45.15	0.56
	RGX	752.77	1.56
	multi-RGX	2648.55	1.90

Table 4.11: Absolute error (AE) of FGX, RGX and multi-RGX in the calculation of the estimated p, μ_H and σ in large dataset with 10000 genes according to the two simulated location-mixture models.

First mixture	True value	FGX	AE	RGX	AE	multi-RGX	AE
p	0.7	1.14	0.63	1.00	0.43	1.14	0.63
μ_H	1	5.65	4.65	5.65	4.65	5.66	4.66
σ	1	0.79	0.21	0.24	0.76	0.23	0.76
Second mixture	True value	FGX	AE	RGX	AE	multi-RGX	AE
p	0.7	1.12	0.60	1.00	0.43	1.12	0.60
μ_H	1	5.61	4.61	5.61	4.61	5.62	4.62
σ	1	0.44	0.56	0.14	0.86	0.13	0.86

both datasets are generated according to the mixture models as described previously. Finally to compare the results, we compute the average absolute error, mean absolute error and root mean square error for estimated S_i in which the first data have 10000 genes, i.e., $i = 1, \ldots, 10000$, and the second data have 20000 genes, i.e., $i = 1, \ldots, 20000$. On the other hand, for the remaining model parameters p, μ_H and σ, we compute directly AE, MAE and RMSE.

Hereby, in the simulation we set the true value of p, μ_H and σ to 0.7, 1 and 1, in order. For the true signals S_i ($i = 1, \ldots, 10000$ and $i = 1, \ldots, 20000$), we initially generate values according to the number of total genes in the range from 2 to 11 from uniform distributions assuming that the genes are measured under 10 distinct concentrations similar to the first two simulated datasets and the intensities are measured on the original scale so that similar to the first two simulations' examples, we aim to evaluate the performance of all three methods under the most problematic range of intensities that is the worse scenario for the comparison.

Table 4.12: Average absolute error (AAE), mean absolute error (MAE) and root mean square error (RMSE) of FGX, RGX and multi-RGX in the calculation of the estimated signals S_i ($i = 1, \ldots, 10000$) in large dataset with 10000 genes according to the two simulated location-mixture models.

First mixture	FGX			RGX			multi-RGX		
	AAE	MAE	RMSE	AAE	MAE	RMSE	AAE	MAE	RMSE
S_i	1.41	8.73	872.85	1.41	8.73	872.85	1.45	8.73	873.41
Second mixture	FGX			RGX			multi-RGX		
	AAE	MAE	RMSE	AAE	MAE	RMSE	AAE	MAE	RMSE
S_i	1.31	8.16	816.43	1.31	8.16	816.43	1.33	8.17	817.08

Then, alike other estimates of other model parameters we calculate the average of all model selection criteria, i.e., AAE, MAE and RMSE, for all signals. Thereby, the average of these criteria for signals is found by using 10000 and 20000 values for the first and second dataset, respectively. Whereas, they are computed for estimated p, μ_H and σ directly. The results are presented in Table 4.11 and 4.12 for the first large dataset (with 10000 genes) whose mixture proportions are arranged according to Equation 4.1 and 4.1, in order. And the outcomes of the second large dataset (with 20000 genes) are listed in Table 4.13 and in Tables 4.14 - 4.15 whose mixture model is generated with respect to Equation 4.1 and 4.1, respectively.

From all findings with large datasets, we observe that the performance of all three indices become very close to each other in such a way that under certain conditions, the findings of multi-RGX are the same with the findings of FGX or RGX or even all the three results become equal to each other. On the other hand, when we observe a difference between the indices, the underlying difference is infinitesimal small meaning that all these alternative methods perform almost equally under large dataset based on the our selected model selection criteria. Moreover, in all models we see that the bias in the estimates decrease considerably, resulting in that the performance of all the three indices are good even under the worse scenario for intensities' level.

Therefore, we conclude that under very large datasets, there is no difference in the application of any suggested model. Whereas, from the results of real and simulated data under small number of genes, we observe that there are differences among all indices. But still there is no unique model which can give the best results according to all model selection criteria. However, we find that FGX, RGX and multi-RGX estimates work well in the bench-mark dataset and for the simulated data with small number of genes, multi-RGX performs better than its strong alternatives, RGX and FGX. Hence, we believe that our suggested method can be seen as a promising approach to estimate the true signals for one-channel microarray datasets.

Table 4.13: Absolute error (AE) of FGX, RGX and multi-RGX in the calculation of the estimated p, μ_H and σ in large dataset with 20000 genes according to the two simulated location-mixture models.

First mixture	True value	FGX	AE	RGX	AE	multi-RGX	AE
p	0.7	1.14	0.63	1.00	0.43	1.14	0.62
μ_H	1	5.64	4.64	5.64	4.64	5.64	4.64
σ	1	1.55	0.55	0.24	0.77	0.24	0.77
Second mixture	True value	FGX	AE	RGX	AE	multi-RGX	AE
p	0.7	1.11	0.59	1.00	0.43	1.11	0.59
μ_H	1	5.67	4.67	5.67	4.67	5.68	4.68
σ	1	0.78	0.22	0.14	0.86	0.14	0.86

Table 4.14: Average absolute error (AAE), mean absolute error (MAE) and root mean square error (RMSE) of FGX and RGX in the calculation of the estimated signals S_i ($i = 1, \ldots, 20000$) in large dataset with 10000 genes according to the two simulated location-mixture models.

First mixture	FGX			RGX		
	AAE	MAE	RMSE	AAE	MAE	RMSE
S_i	1.41	8.69	1229.42	1.41	8.69	1229.42
Second mixture	FGX			RGX		
	AAE	MAE	RMSE	AAE	MAE	RMSE
S_i	1.32	8.20	1159.14	1.32	8.20	1159.14

Table 4.15: Average absolute error (AAE), mean absolute error (MAE) and root mean square error (RMSE) of multi-RGX in the calculation of the estimated signals S_i ($i = 1, \ldots, 20000$) in large dataset with 10000 genes according to the two simulated location-mixture models.

First mixture	multi-RGX		
	AAE	MAE	RMSE
S_i	1.45	8.70	1230.33
Second mixture	multi-RGX		
	AAE	MAE	RMSE
S_i	1.34	8.20	1160.36

CHAPTER 5

CONCLUSION AND OUTLOOK

This thesis gives a deep introduction to the idea of oligonucleotide and microarray. Thereby, it gives the information about arrays, namely, chips, and on these arrays, we can display the known segments of particular gene sequences, which is called probes. On these probes, there are two components; the perfect match (PM), which is the perfect transcription of the cRNA and the mismatch (MM) that is used to measure the faulty signals on the arrays by changing the 13th base pair of the PM.

In Chapter 1, we have explained the three main sources of variation of signals, listed as nonspecific hybridization, background signal and stray signal. These three types of errors can be defined as the systematic error.

Then, we have discussed the possible ways to normalize data in order to discard the systematic error in the measurements. As explained in details in Chapter 2, there are two major approaches of normalization for the microarray dataset. The first approach is based on the ANOVA idea where the normalization is implemented while the analysis of the data is performed. As presented in the associated part in junction with the current study about this approach such as robust estimation procedure via MMLE and AMMLE, this method can be computationally demanding if the inference is done via the least square or MLE methods. Whereas, as the second approach, the normalization can be also performed in advance of the actual analysis by eliminating possible sources of noisy signals in a particular order. In this thesis, we have preferred this second approach since it is shown that this is computationally less costly.

Later, we have introduced the term gene expression index as the method or model applied to estimate the true expression level in the oligonucleotides, i.e., one-channel microarray. For this purpose, we have given information about the most popular indices such as MAS 5.0, MBEI, RMA, mgMOS, GC-RMA, BGX, multi-mgMOS, FGX and RGX, and indicated that they have both advantages and disadvantages in their perspectives.

Since we have stated these methods have some problems besides their advantages, we have proposed a new model, which is called multi-RGX, in order to handle the disadvantages of recently developed methods, which are FGX and RGX.

Our method, namely, multi-RGX, works on the gene and probe specific signal level in the measurement of microarray. We have generated explicit expressions for each model parameter by the help of the modified maximum likelihood method. Also, we have developed explicit forms of the variances and covariances of model parameters by applying observed Fisher

Information Matrix. Apart from our suggested approach, we have also constructed other alternative models, which accept more flexible assumptions, whereas, do not have either close form for the expression of estimators or are not defined under feasible region.

In the assessment of our proposal gene expression, we have analyzed two real datasets that are from Affymetrix and GeneLogic platforms. These measurements are benchmarks data in the comparison of various methods and include spike-in genes under different concentrations. We evaluate the performance of our model in terms of accuracy, linearity with respect to the changes in concentrations and computational demand. Then, we have further assessed the outcomes of our approach in simulated datasets. In the comparison, we have selected the strongest alternatives of multi-RGX, which are FGX and RGX functions. In the analyses, we have compared the accuracy and computational demand of each approach with certain criteria. For the accuracy, the evaluation has been conducted under the absolute error, mean square error and root mean square error values. On the other hand, for the speed of the function, the assessment has been done via the real and CPU time. According to the results presented in Chapter 4, we have observed that multi-RGX can keep high accuracy regarding alternatives gene expression indices while preserving less computational cost under small dataset. But there is no significant difference among any models in large dataset with ten thousands of genes.

Moreover, as this index can use both probe and gene level of observations and has explicit forms for each estimator with their associated covariance and variance terms, its advantages can be better observed in small or moderately large data. Thus, we believe that multi-RGX can be seen as a promising model in order to infer the true signal from one-channel microarrays.

As the extension of this study, we can implement adaptive modified maximum likelihood approach in the estimation of model parameters (Ülgen, 2010; Dönmez, 2010; Tiku and Sürücü, 2009) in order to increase the accuracy of the estimators. Furthermore, as a nonlinear relationship is observed between signal and concentrations in particular under low and high concentrations, we consider to implement multiple adaptive regression splines (MARS) approach in order to model this relationship via partial linear functions. This idea has been previously performed in the study of Xu et al. (2010) in the analysis of the chip sequence data to detect the function of histone modification levels and investigating the gene relationship between chromatin feature levels and gene expression again in chip sequence data (Dong et al., 2012). Moreover, Chang et al. (2008) apply this method in order to find out the cut-off point for intensity in a microarray study so that the conserved and divergent genes in bacterial identification and characterization can be detected. Furthermore, this approach is already performed in the analysis of different multiple nonlinear datasets from financial to environmental studies (Alp et al., 2011; Taylan et al., 2007). Accordingly, if we adapt this technique in our model by fitting distinct functions for low, medium and high concentration data, separately, we can explain the behaviour between signal and concentration level better than a single linear model fitted to the whole dataset.

CHAPTER 6

REFERENCES

1. Affymetrix (2002). Statistical algorithms description document. Affymetrix, 1-28.

2. Akkaya, A. D. and Tiku, M. L. (2007). Robust estimation in multiple linear regression model with non-gaussian noise. Automatica, 44, 407-417.

3. Asbby, R. S. and Feldkaemper, M. P. (2010). Gene expression within the amacrine cell layer of chicks after myopic and hyperopic defocus. Investigative Ophthalmology and Visual Science, 51 (7), 3726-3735.

4. Augugliaro, L. and Mineo, A. M. (2010). A Statistical Calibration Model for Affymetrix Probe Level Data. 121-128. Chapter in: Data Analysis and Classification Studies in Classification, Data Analysis, and Knowledge Organization, Springer.

5. Bhattacharyya, G. K. (1985). The asymptotic of maximum likelihood and related estimators on type two censored data. Journal of American Statistical Association, 80, 398-404.

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

7. Bonner-Weir, S., Sharma, A. and Aguayo-Mazzucato, C. (2012). Compositions and methods for promoting beta cell maturity. Joslin Diabeters Center, Inc.

8. Buler, M., Aatsinki, S. M., Skoumal, R., Komka, Z., Toth, M., Kerkela, R., Georgiadi, A., Kersten, S. and Hakkola, J. (2011). Energy sensing factors coactivator pgc-1a and amp-activated protein kinase control expression of inflammatory mediators in liver: Induction of interleukin 1 receptor antagonist. Journal of Biological Chemistry, 287 (3), 1847-60.

9. Chang, C. W., Zou, W. and Chen, J. J. (2008). A new method for gene identification in comparative genomic analysis. Journal of Data Science, 6, 415-427.

10. Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. and Speed, T. P. (2003). A benchmark for Affymetrix GeneChip expression measures. Bioinformatics, 1 (1), 1-10.

11. Cusumano, C. K., Hung, C. S., Chen, S. L. and Hultgren, S. J. (2010). Virulence plasmid harbored by uropathogenic escherichia coli functions in acute stages of pathogenesis. Infection and Immunity, 78(4), 1457-1467.

12. Dalby, A. R., Emam, I. and Franke, R. (2012). Analysis of gene expression data from non-small cell lung carcinoma cell lines reveals distinct sub-classes from those identified at the phenotype level. PLOS ONE, 7(11), doi: 10.1371/journal.pone.0050253, 1-13.

13. Dash, S., Hemert, J. W., Hong, L., Wise, R. P. and Dickerson, J. A. (2012). Plexdb: Gene expression resources for plants and plant pathogens. Nucleic Acids Research, 40, doi: 10.1093/nar/gkr938, 1-8.

14. Derrien, M., Baarlen, P. V., Hooiveld, G., Norin, E., Muller, M. and Vos, W. M. (2011). Modulation of mucosal immune response, tolerance, and proliferation in mice colonized by the mucin-degrader akkermansia muciniphila. Frontiers in Microbiology, 2, doi: 10.3389/fmicb.2011.00166, 1-14.

15. Dong , X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigo, R., Birney, E. and Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. Genome Biology, 13 (R53), doi: 10.1186/gb-2012-13-9-r53, 1-17.

16. Dönmez, A. (2010). Adaptive estimation and hypothesis testing methods. Ph.D Thesis, Middle East Technical University, Ankara.

17. Dunnett, C. W. (1982). Robust multiple comparisons. Communication in Statistics: Theory and Methods, 11 (22), 2611-2629.

18. Grant, B., Nickel, B. and Vershon, A. (2013). The Basics: DNA, RNA, proteins, transcription and translation. Introduction to Molecular Biology and Biochemistry Research Lecture Notes, Department of Molecular Biology and Chemistry, The Rutgers University, New Jersey, U.S.A, 1-12.

19. Greco, D., Vellonen, K. S., Turner, H. C., Hakli, M., Tervo, T., Auvinen, P., Wolosin, J. M. and Urtti, A. (2010). Gene expression analysis in sv-40 immortalized human corneal epithelial cells cultured with an air-liquid interface. Molecular Vision, 16, 2109-2120.

20. Guerri, E. C., Garcia, O. S., Pascual, V. F., Gerboles, P. J. and Guillermo, A. L. (2012). Method and kit for the prognosis of mantle cell lymphoma. United States Patent Application Publication.

21. Ha, M., Ng, D. W., Li, W. and Chen, Z. J. (2011). Coordinated histone modifications are associated with gene expression variation within and between species. Genome Research, 21, 590-598.

22. Hein, A. K., Richardson, S., Causton, H. C., Ambler, G. K. and Green, P. J. (2005). BGX: a fully bayesian gene expression index for Affymetrix genechip data. Biostatistics, 6 (3), 349-373.

23. Herlofsen, S. R., Kuchler, A. M., Melvik, J. E. and Brinchmann, J. E. (2011). Chondrogenic differentiation of human bone marrow-derived mesenchymal stem cells in self-gelling alginate discs reveals novel chondrogenic signature gene clusters. Tissue Engineering, 17 (7 and 8), 1003-1013.

24. Hubbell, E., Liu, W. and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics, 18 (12), 1585-1592.

25. Irizzary, R. A., Hobbs, B., Collin, F. and Speed, T. P. (2003). *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*. Biostatistics, 249-264.

26. Irizarry, R. A. and Zillio, M. J. (2010). Gene expression barcode for normal and diseased tissue classification. John Hopkins University.

27. Jailwala, P., Waukau, J., Glisic, S., Jana, S., Ehlenbach, S., Hessner, M., Alemzadeh, R., Matsuyama, S., Laud, P., Wang, X. and Ghosh, S. (2009). Apoptosis of cd4 cd25high t cells in type 1 diabetes may be partially mediated by il-2 deprivation. PLOS One, 4 (8), e65277/1-13.

28. Janne, P. A., Engelman, J. and Cantley, L. C. (2012). Methods for treating cancer resistant to erbb therapeutics. Dana Farber Cancer Institute Inc.

29. Kartal-Koç, E, Batmaz, İ. and Weber G. W. (2012). Robust regression metamodelling of complex systems: the case of solid rocket motor performance metamodelling. 221-251. Chapter in: Advances in Intelligent Modelling and Simulation: Simulation Tools and Applications, Springer-Verlag, Berlin Heidelberg.

30. Kennedy, R. E. (2008). Probe Level Analysis of Affymetrix Microarray Data. Ph.D. Thesis, University of Mississippi, Mississippi.

31. Kerber, R. A., O'Brien, E. and Cawthon, R. M. (2009). Gene expression profiles associated with aging and mortality in humans. Aging Cell, 8, 239-250.

32. Kerr, M. K., Martin, M. and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. Journal of Computational Biology, 7 (6), 819-837.

33. Kerr, M. K. and Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. Genetical Research, 77, 123-128.

34. Kerr, M. K. (2009). Computational analysis of gene expression data. Ph. D. Thesis, Dublin City University, Dublin.

35. Kliebenstein, D. J., West, M. A. L., Leeuwen, H., Kim, K., Doerge, R. W., Michelmore, R. W. and St. Clair, D. A. (2006). Genomic survey of gene expression diversity in arabidopsis thaliana. Genetics, 172, 1179-1189.

36. Kostek, M. C., Chen, Y. W., Cuthbertson, D. J., Shi, R., Fedele, M. J., Esser, K. A. and Rennie, M. J. (2007). Gene expression responses over 24 h to lengthening and shortening contractions in human muscle: Major changes in csrp3, mustn1, six1, and fbxo32. Physiol Genomics, 31, 42-52.

37. Le, D. T., Nishiyama, R., Watanabe, Y., Tanaka , M., Seki, M., Ham , L. H., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2012). Differential gene expression in soybean leaf tissues at late developmental stages under drought stress revealed by genome-wide transcriptome analysis. PLOS One, 7 (11), e49522/1-10.

38. Lemon, W. J., Palatini, J. J. T., Krahe, R. and Wright, F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. Bioinformatics, 18, 1470-1476.

39. Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays:expression index computation and outlier detection. PNAS, 98, 31–36.

40. Meibner, T., Seckinger, A., Reme, T., Hielscher, T., Mohler, T., Neben, K., Goldschmidth, H., Klein, B. and Hose, D. (2011). Gene expression profiling in multiple myeloma reporting of entities, risk, and targets in clinical routine. Clinical Cancer Research, 17 (23), 7240-7247.

41. Michael, T. P., Breton, G., Hazen, S. P., Priest, H., Mockler, T. C., Kay, S. A. and Chory, J. (2008). A morning-specific phytohormone gene expression program underlying rhythmic plant growth. PLOS Biology, 6 (9), 1887-1898.

42. Milo, M., Fazelit, A., Niranjan, M. and Lawrence, N. D. (2003). A probabilistic model for the extraction of expression levels from oligonucleotide arrays. Biochemical Society Transactions, 31, 1510-1512.

43. Mitchell, S., Ota, A., Foster, W., Zhang, B., Fang, Z., Patel, S., Nelson, S. F., Horwath, S. and Wang, Y. (2005). Distinct gene expression profiles in adult mouse heart following targeted map kinase activation. Physiol Genomics, 25, 50-59.

44. Noyes, H., Brass, A., Obara, I., Anderson, S., Archibald, A. L., Bradley, D. G., Fisher, P., Freeman, A., Gibson, J., Gicheru, M., Hall, L., Hanotte, O., Hulme, H., McKeever, D., Murray, C., Oh, S. J., Tate, C., Smith, K., Tapio, M., Wambugu, J., Williams, D. J., Agaba, M. and Kemp, S. J. (2011). Genetic and expression analysis of cattle identifies candidate genes in pathways responding to trypanosoma congolense infection. PNAS, 108 (22), 9304-9309.

45. Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K. and Ohta, H. (2007). ATTED-II: A database of co-expressed genes and cis elements for identifying co-regulated gene groups in arabidopsis, 35, 2863-2869.

46. Purutçuoğlu, V. (2007). Bayesian methods for gene network analysis. Ph. D. Thesis, Lancaster University, Lancaster.

47. Purutçuoğlu, V. (2012). Robust gene expression index. Mathematical Problems in Engineering, doi: 10.1155/2011/182758, 1-17.

48. Purutçuoğlu, V. and Wit, E. (2006). FGX: a frequentist gene expression index for Affymetrix arrays. Biostatistics, 433-437.

49. Purutçuoğlu, V. and Wit, E. (2008). Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis, 3 (4), 851-886.

50. Purutçuoğlu, V., Kayış, E. and Weber, G. W. (2012). Survey of background normalizations for Affymetrix arrays and a case study. 199-219. Chapter in: Advances in Intelligent Modelling and Simulation: Simulation Tools and Applications, Springer-Verlag, Berlin Heidelberg.

51. Reppe, S., Stilgren, L., Abrahamsen, B., Olstad, O. K., Cero, F., Brixen, K., Nissen-Meyer, L. S. and Gautvik, K. M. (2007). Abnormal muscle and hematopoietic gene expression may be important for clinical morbidity in primary hyperparathyroidism. American Journal of Physiology - Endocrinology and Metabolism, 292, E1465-E1473.

52. Sanchez, A. and Ruiz de Villa, M. C. (2008). A tutorial review of microarray data analysis. Universitat de Barcelona, Spain.

53. Sarmah, C. K. and Samarasinghe, S. (2011). Microarray gene expression: a study of between-platform. Computers in Biology and Medicine, 41 (10), 980-986.

54. Alp, Ö. S., Büyükbebeci, E., Iscanoglu Cekic, A., Yerlikaya Özkurt, F., Taylan, P. and Weber, G. W. (2011). CMARS and GAM and CQP - modern optimization methods applied to international credit default prediction. Journal of Computational and Applied Mathematics, 235, 4639-4651.

55. Schippert, R., Schaeffel, F. and Feldkaemper, M. P. (2008). Microarray analysis of retinal gene expression in chicks during imposed myopic defocus. Molecular Vision, 14, 1589-1599.

56. Shlien, A. M. and Malkin, D. D. (2010). Methods of fetermining risk for cancer. Toronto, CA.

57. Staudt, L. M., Wright, G. W., Dave, S. and Tan, B. K. (2011). Methods for identifying, diagnosing, and predicting survival of lymphomas. Government of the USA, Department of Health and Human Services.

58. Steen, K. (2002). A Biologist's Guide to Analysis of DNA Microarray Data. John Wiley Sons Limited.

59. Stekel, D. (2003). Microarray Bioinformatics. Cambridge University.

60. Taylan, P., Weber, G. W. and Beck, A. (2007). New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and techology. Optimization, 56 (5 and 6), 675-698.

61. Tiku, M. L., Tan, W. Y. and Balakrishnan, N. (1986). Robust inference. Marcel Dekker, New York.

62. Tiku, M. L. and Akkaya, A. D. (2004). Robust estimation and hypothesis testing. New Age International Limited Publishers, New Delhi.

63. Tiku, M. L. and Sürücü, B. (2009). MMLEs are as good as M-estimators or better. Statistics and Probability Letters, 79, 984-989.

64. Ülgen, E. B. (2010). Robust estimation and hypothesis testing in microarray analysis. Ph.D Thesis, Middle East Technical University, Ankara.

65. Vaughan, D. C. (1992). On the Tiku-Suresh method of estimation. Communications in Statistics: Theory and Methods, 21, 451-469.

66. Venezia, T. A., Merchant, A. A., Ramos, C. A., Whitehouse, N. L., Young, A. S., Shaw, C. A. and Goodell, M. A. (2004). Molecular signatures of proliferation and quiescence in hematopoietic stem cells. PLOS Biology, 2 (10), 1640-1651.

67. Wit, E. C. and McClure J. (2004). Statistics for Microarrays: Design, Analysis and Inference. John Wiley Sons Limited.

68. Wu, Z., Irizzary, R. A., Gentleman, R., Martinez-Murillo, F. and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. Journal of American Statistical Association, 99 (468), 909-917.

69. Yang, X., Regan, K., Huang, Y., Zhang, Q., Li, J., Seiwert, T. Y., Cohen, E. E. W., Xing, H. R. and Lussier, Y. A. (2012). Single sample expression-anchored mechanisms predict survival in head and neck cancer. Computational Biology, 8 (1), 1-18.

70. Xu, X., Hoang, S., Mayo, M. W. and Bekiranov, S. (2010). Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. BMC Bioinformatics, 11 (396), 1-20.

APPENDIX A

DERIVATION of the multi-RGX ESTIMATORS

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we consider the following distributional assumption on the log-scale:

$$\mathrm{PM}_{ij} = a_{ij} \sim \mathrm{LTS}(S_i + \mu_H, \sigma^2)$$

and

$$\mathrm{MM}_{ij} = b_{ij} \sim \mathrm{LTS}(pS_i + \mu_H, \sigma^2), \tag{A.1}$$

where LTS denotes the long-tailed symmetric density.

Thereby, the corresponding likelihood is found via:

$$\begin{aligned} L(S_i, \mu_H, p \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a_{ij}-S_i-\mu_H)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(b_{ij}-pS_i-\mu_H)^2}{2\sigma^2}}, \end{aligned}$$

which is proportional to

$$L \propto (\frac{1}{\sigma}) \prod_{i=1}^{n}\prod_{j=1}^{m} (1 + \frac{z_{a_{ij}}^2}{k})^{-v} (\frac{1}{\sigma}) \prod_{i=1}^{n}\prod_{j=1}^{m} (1 + \frac{z_{b_{ij}}^2}{k})^{-v},$$

where $v \geq 2$, $k = 2v - 3$, and

$$\begin{aligned} z_{a_{ij}} &= \frac{(a_{ij} - S_i - \mu_H)}{\sigma} \\ z_{b_{ij}} &= \frac{(b_{ij} - pS_i - \mu_H)}{\sigma} \\ v &\geq 2 \\ k &= 2v - 3. \end{aligned}$$

Then, the logarithm of L is found as

$$\begin{aligned}\ln L \quad &\propto \quad -2nm\ln\sigma + \sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[1+\frac{a_{ij}-S_i-\mu_H}{k\sigma^2}\right]^{-v} \\ &+ \quad \sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[1+\frac{b_{ij}-pS_i-\mu_H}{k\sigma^2}\right]^{-v}.\end{aligned}$$

In order to calculate the MLE of the model parameters, the first derivatives of the function $\ln L$ are taken with respect to each parameter as below:

$$\begin{aligned}\frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma}}{\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2}} + \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\frac{(b_{ij}-pS_i-\mu_H)}{\sigma}}{\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}} \\ \frac{\partial \ln L}{\partial p} &= \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\frac{S_i(b_{ij}-pS_i-\mu_H)}{\sigma}}{\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}} \\ \frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma}}{\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2}} + \frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{-p(b_{ij}-pS_i-\mu_H)}{\sigma}}{\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}} \\ \frac{\partial \ln L}{\partial \sigma} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma}+\frac{(a_{ij}-S_i-\mu_H)}{\sigma^3(1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2})}\frac{2v}{k}\right] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma}+\frac{(b_{ij}-pS_i-\mu_H)}{\sigma^3(1+\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2})}\frac{2v}{k}\right],\end{aligned}$$

in which

$$z_{a_{ij}} = \frac{(a_{ij}-S_i-\mu_H)}{\sigma} \text{ and } z_{b_{ij}} = \frac{(b_{ij}-pS_i-\mu_H)}{\sigma}.$$

Then, by approximating the common nonlinear functions $g(z) = \frac{z}{1+\frac{z^2}{k}}$ for PM:

$$\text{as } g(z_{a_{ij}}) = \frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma}}{1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma^2}}$$

and for MM:

$$g(z_{b_{ij}}) = \frac{\frac{(b_{ij}-pS_i-\mu_H)}{\sigma}}{1+\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma^2}},$$

their first order Taylor expansions can be derived by:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

where

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2}.$$

In these expressions, $z_{a_{i(j)}}$ and $z_{b_{i(j)}}$ show the ordered probes (in increasing magnitude) for each gene i in PM and MM standardized intensities, respectively. Accordingly, $g(z_{a_{i(j)}})$ and $g(z_{b_{i(j)}})$ are their associated linearized function via the Taylor series. Thereby, the partial derivatives of MLE are as follows:

$$\frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} g(z_{b_{i(j)}}) \tag{A.2}$$

$$\frac{\partial \ln L}{\partial p} = \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} S_i g(z_{b_{i(j)}}) \tag{A.3}$$

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k\sigma}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2vp}{k\sigma}\sum_{j=1}^{m} g(z_{b_{i(j)}}) \tag{A.4}$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{2nm}{\sigma} + \frac{2v}{k\sigma}\sum_{j=1}^{m} g(z_{a_{i(j)}}) z_{a_{i(j)}} + \frac{2v}{k\sigma}\sum_{j=1}^{m} g(z_{b_{i(j)}}) z_{b_{i(j)}}. \tag{A.5}$$

Finally, by setting the Equations (A.2) - (A.5) to zero, we get the MML estimates of parameters. Accordingly, from Equation (A.2):

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} g(z_{b_{i(j)}}) &= 0 \\
\frac{1}{\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) &= 0 \\
\sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha_j + \beta_j z_{a_{i(j)}} + \alpha_j + \beta_j z_{b_{i(j)}}) &= 0 \\
\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j z_{a_{i(j)}} + \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j z_{b_{i(j)}} &= 0,
\end{aligned}$$

by taking $\sum_{j=1}^{m}\alpha_j = 0$ because of the symmetry.

Then,

$$\frac{\partial \ln L}{\partial \mu_H} = \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j z_{a_{i(j)}} + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j z_{b_{i(j)}} = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j \frac{(a_{i(j)} - S_i - \mu_H)}{\sigma} + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j \frac{(b_{i(j)} - pS_i - \mu_H)}{\sigma} = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j (a_{i(j)} - S_i - \mu_H) + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j (b_{i(j)} - pS_i - \mu_H) = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i = -\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j b_{i(j)} + p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i + 2\mu_H n\sum_{j=1}^{m}\beta_j$$

$$2n\mu_H \sum_{j=1}^{m}\beta_j = \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j (a_{i(j)} + b_{i(j)}) - (p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i,$$

when $a_{i(j)}$ and $b_{i(j)}$ display the ordered probes of PM and MM, respectively, for gene i ($i = 1, \ldots, n$). As a result, $\hat{\mu}_H$ is found as:

$$\hat{\mu}_H = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j (a_{i(j)} + b_{i(j)}) - (p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i}{2n\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j}. \tag{A.6}$$

On the other side, from Equation (A.4):

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma} g(z_{a_{i(j)}}) + \frac{2vp}{k}\sum_{j=1}^{m}\frac{1}{\sigma} g(z_{b_{i(j)}}) = 0.$$

So

$$\frac{2v}{k}\frac{1}{\sigma}\left[\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + pg(z_{b_{i(j)}}))\right] = 0$$

$$\frac{2v}{k\sigma}\sum_{j=1}^{m}((\alpha_j + \beta_j z_{a_{i(j)}}) + \frac{2v}{k\sigma}\sum_{j=1}^{m}((\alpha_j + \beta_j z_{b_{i(j)}}) = 0$$

$$\frac{1}{\sigma^2}\sum_{j=1}^{m}\beta_j (a_{i(j)} - S_i - \mu_H) + \frac{p}{\sigma^2}\sum_{j=1}^{m}\beta_j (b_{i(j)} - pS_i - \mu_H) = 0$$

$$\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j S_i - \mu_H\sum_{j=1}^{m}\beta_j + p\sum_{j=1}^{m}\beta_j b_{i(j)} - p^2\sum_{j=1}^{m}\beta_j S_i - p\mu_H\sum_{j=1}^{m}\beta_j = 0$$

$$(p^2+1)S_i\sum_{j=1}^{m}\beta_j = \sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_H\sum_{j=1}^{m}\beta_j = 0\ ,$$

we obtain the estimate of the true signal for each gene i as below:

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_H \sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j} . \tag{A.7}$$

However, by substituting A.7 into A.6, we can define $\hat{\mu}_H$ in an alternative way as follows:

$$\begin{aligned}
\hat{\mu}_H &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)})}{2n\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j} \\
&- \frac{(p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[\frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_H\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j}\right]}{2n\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j} \\
\hat{\mu}_H &= \frac{p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j b_{(ij)}}{n(p-1)\sum_{j=1}^{m}\beta_j}.
\end{aligned} \tag{A.8}$$

Similarly, by substituting A.8 into A.7, an alternative form of $\hat{S}_i$ can be written as below:

$$\begin{aligned}
\hat{S}_i &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_H\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j} \\
&= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p^2+1)\sum_{j=1}^{m}\beta_j} \\
&- \frac{(p+1)\left[\frac{p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j b_{(ij)}}{n(p-1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j}
\end{aligned}$$

Accordingly,

$$\begin{aligned}
\hat{S}_i &= \frac{1}{(p^2+1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j} + \frac{p}{(p^2+1)}\frac{\sum_{j=1}^{m}\beta_j b_{i(j)}}{\sum_{j=1}^{m}\beta_j} \\
&- \frac{p(p+1)}{n(p-1)(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j a_{(ij)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j} \\
&+ \frac{(p+1)}{n(p-1)(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j b_{(ij)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j}.
\end{aligned}$$

Let

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j a_{i(j)}}{n} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\beta_j a_{i(j)}}{n} = \sum_{j=1}^{m}\beta_j\frac{\sum_{i=1}^{n}a_{i(j)}}{n} = \sum_{j=1}^{m}\beta_j\bar{a}_{.j},$$

where $\bar{a}_{.j} = \sum_{i=1}^{n} \frac{a_{i(j)}}{n}$ for total number of genes. Another form of $\hat{S}_i$ can be described in the following way:

$$\begin{aligned}\hat{S}_i &= \frac{(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})p^2}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)} \\ &+ \frac{(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j} + \sum_{j=1}^{m}\beta_j \bar{b}_{.j})p}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)} \\ &+ \frac{\sum_{j=1}^{m}\beta_j \bar{b}_{.j} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j}, \end{aligned} \tag{A.9}$$

for $\bar{b}_{.j} = \sum_{i=1}^{n} \frac{a_{i(j)}}{n}$.

Similarly, $\hat{\mu}_H$ can also be shown by:

$$\hat{\mu}_H = \frac{p\sum_{j=1}^{m}\beta_j \bar{a}_{.j} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j}}{(p-1)\sum_{j=1}^{m}\beta_j}. \tag{A.10}$$

On the other hand, from Equation (A.3), the estimate of the common fraction term p can be found as below:

$$\begin{aligned}\frac{\partial \ln L}{\partial p} = \frac{2v}{k\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m} S_i g(z_{b_{i(j)}}) &= 0 \\ \sum_{i=1}^{n}\sum_{j=1}^{m} S_i g(z_{b_{i(j)}}) &= 0 \\ \sum_{i=1}^{n}\sum_{j=1}^{m} S_i \alpha_j + \sum_{j=1}^{m}\beta_j \frac{(b_{i(j)} - pS_i - \mu_H)}{\sigma} S_i &= 0 \\ \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i b_{i(j)} - p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i^2 - \mu_H \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j S_i &= 0. \end{aligned}$$

Then, by substituting Equations (A.9) and (A.10) into the equation,

$$
\begin{aligned}
0 &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j b_{i(j)}\left\{\left[\frac{(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})p^2}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)}\right]\right. \\
&+ \left[\frac{(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j} + \sum_{j=1}^{m}\beta_j \bar{b}_{.j})p}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)}\right] \\
&+ \left.\left[\frac{\sum_{j=1}^{m}\beta_j \bar{b}_{.j} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j}\right]\right\} \\
&- p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left(\left[\frac{(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})p^2}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)}\right]\right. \\
&+ \left[\frac{(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j} + \sum_{j=1}^{m}\beta_j \bar{b}_{.j})p}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)}\right] \\
&+ \left.\left[\frac{\sum_{j=1}^{m}\beta_j \bar{b}_{.j} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j}\right]\right)^2 \\
&- \left[\frac{p\sum_{j=1}^{m}\beta_j \bar{a}_{.j} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j}}{(p-1)\sum_{j=1}^{m}\beta_j}\right] \\
&\times \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left(\left[\frac{(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})p^2}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)}\right]\right. \\
&+ \left[\frac{(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j} + \sum_{j=1}^{m}\beta_j \bar{b}_{.j})p}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j)}\right] \\
&+ \left.\left[\frac{\sum_{j=1}^{m}\beta_j \bar{b}_{.j} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p-1)(p^2+1)(\sum_{j=1}^{m}\beta_j}\right]\right).
\end{aligned}
$$

By equating the above equation to 0 and solving in terms of p, we get the following roots for $\hat{p}$:

$$
\begin{aligned}
\hat{p}_1 &= \hat{p}_2 = 1 \\
\hat{p}_3 &= \frac{(SS_b - SS_a) - \sqrt{(SS_a - SS_b)^2 + 4SS_{ab}^2}}{2SS_{ab}} \\
\hat{p}_4 &= \frac{(SS_b - SS_a) + \sqrt{(SS_a - SS_b)^2 + 4SS_{ab}^2}}{2SS_{ab}},
\end{aligned}
$$

where

$$
\begin{aligned}
SS_a &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j}\right]^2 \\
SS_b &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j}\right]^2 \\
SS_{ab} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\left[(\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{a}_{.j})(\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j \bar{b}_{.j})\right].
\end{aligned}
$$

Finally, from Equation (A.5), we can get MLE of σ by setting this expression to zero via

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \sigma} &= -\frac{2m}{\sigma} + \frac{2v}{k\sigma^2}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma^2}\sum_{j=1}^{m} g(z_{b_{i(j)}}) \\
&= -\frac{2m}{\sigma} + \frac{2v}{k\sigma^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) = 0 \\
\frac{2m}{\sigma} &= \frac{2v}{k\sigma^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\sigma &= \frac{v}{knm}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})).
\end{aligned}
$$

Accordingly,

$$
\begin{aligned}
\frac{v}{knm} &= \frac{v}{knm}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j a_{i(j)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j S_i - \mu_H\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j\right] \\
&+ \frac{v}{knm}\left[\frac{1}{\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)} - S_i - \mu_H)^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j b_{i(j)}\right] \\
&- \frac{v}{knm}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j S_i p + \mu_H\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j\right] \\
&+ \frac{v}{knm}\left[\frac{1}{\sigma}\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)} - S_i - \mu_H)^2\right].
\end{aligned}
$$

Thus,

$$\begin{aligned}
\sigma^2 &= \frac{\sigma v}{knm}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j(a_{i(j)}-b_{i(j)})-(p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j s_i\right] \\
&+ \frac{\sigma v}{knm}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j(a_{i(j)}-b_{i(j)})-(p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j S_i\right], \\
nm\sigma^2 &= \frac{\sigma v}{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j(a_{i(j)}-b_{i(j)})\right] \\
&- \frac{v}{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)}-S_i-\mu_H)^2+\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)}-S_i-\mu_H)^2\right].
\end{aligned}$$

Hence, the MML estimate for σ can be presented by:

$$\hat{\sigma} = \frac{B+\sqrt{B^2+4nmC}}{nm},$$

for S_i $(i = 1,\ldots,n)$. By adjusting the degree of freedom as $2nm-(n+2)$, the estimate of σ can be indicated as follows:

$$\hat{\sigma} = \frac{B+\sqrt{B^2+4nmC}}{nm-n-2},$$

in which

$$\begin{aligned}
B &= \frac{v}{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j(a_{i(j)}-b_{i(j)})\right], \\
C &= \frac{v}{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(a_{i(j)}-\hat{S}_i-\hat{\mu}_H)^2+\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(b_{i(j)}-\hat{p}\hat{S}_i-\hat{\mu}_H)^2\right].
\end{aligned}$$

A.1 Observed Fisher Information Matrix and Estimators for Variances and Covariances

The Fisher Information Matrix is generated by making use of the second partial derivatives of the loglikelihood function with respect to each parameter as below:

$$
\begin{aligned}
I_{11} &= -\frac{\partial^2 l}{\partial {\mu^2}_H} \\
&= \frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right] \\
I_{22} &= -\frac{\partial^2 l}{\partial p^2} \\
&= \frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}S_i^2\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2} \\
I_{ii} &= -\frac{\partial^2 l}{\partial S_i^2} \\
&= \frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2} \\
I_{12} &= I_{21} = -\frac{\partial^2 l}{\partial \mu_H \partial p} = \frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}S_i\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
I_{1i} &= I_{i1} = -\frac{\partial^2 l}{\partial S_i \partial \mu_H} \\
&= \frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2} \\
I_{2i} &= I_{i2} = -\frac{\partial^2 l}{\partial S_i \partial p} \\
&= -\frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}+\frac{2vp}{k\sigma^2}\sum_{j=1}^{m}S_i\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2} \\
I_{ik} &= I_{ki} = -\frac{\partial^2 l}{\partial S_i \partial S_k} = 0,
\end{aligned}
$$

where $i, k = 1, \ldots, n$ present the gene and $j = 1, \ldots, m$ shows the probe indicator, respectively. Moreover, l stands for $\ln L$, the loglikelihood function.

Then, the Fisher information matrix is derived as below:

$$I = \begin{bmatrix} \frac{\partial^2 l}{\partial \mu^2{}_H} & \frac{\partial^2 l}{\partial \mu_H \partial p} & \frac{\partial^2 l}{\partial \mu_H \partial S_1} & \cdots & \frac{\partial^2 l}{\partial \mu_H \partial S_n} \\ \frac{\partial^2 l}{\partial p \partial \mu_H} & \frac{\partial^2 l}{\partial p^2} & \frac{\partial^2 l}{\partial p \partial S_1} & \cdots & \frac{\partial^2 l}{\partial p \partial S_n} \\ \frac{\partial^2 l}{\partial S_1 \partial \mu_H} & \frac{\partial^2 l}{\partial S_1 \partial p} & \frac{\partial^2 l}{\partial S_1^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 l}{\partial S_n \partial \mu_H} & \frac{\partial^2 l}{\partial S_n \partial p} & 0 & \cdots & \frac{\partial^2 l}{\partial S_n^2}. \end{bmatrix}$$

In order to find the variance-covariance matrix we take the advantage of the above information matrix. Hereby, we reformulate the matrix as follows:

$$I = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where A is a (2×2) submatrix containing the entries at the top left hand side of I, B represents the $(2 \times n)$ submatrix at the top of right hand side of I. Finally, C indicates the $(n \times n)$ diagonal submatrix at the bottom right hand side of I.

We can represent the inverse of the matrix I as below:

$$I^{-1} = \begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix},$$

where

$$P = (A - BC^{-1}B^T)^{-1} \tag{A.11}$$
$$Q = -(C^{-1}B^T P)^T \tag{A.12}$$
$$R = C^{-1} - C^{-1}B^T Q. \tag{A.13}$$

From Equation (A.11), P is found via:

$$P = [A - B]^{-1},$$

whose entries are given below:

$$A = \begin{bmatrix} \frac{\partial^2 l}{\partial \mu^2{}_H} & \frac{\partial^2 l}{\partial \mu_H \partial p} \\ \frac{\partial^2 l}{\partial p \partial \mu_H} & \frac{\partial^2 l}{\partial p^2} \end{bmatrix}$$

and

$$B = \begin{bmatrix} \sum \frac{(\partial^2 l/\partial S_i \partial \mu_H)^2}{-\partial^2 l/\partial S_i^2} & \sum \frac{(-\partial^2 l/\partial S_i \partial \mu_H)(-\partial^2 l/\partial S_i \partial p)}{-\partial^2 l/\partial S_i^2} \\ \sum \frac{(-\partial^2 l/\partial S_i \partial \mu_H)(-\partial^2 l/\partial S_i \partial p)}{-\partial^2 l/\partial S_i^2} & \sum \frac{(\partial^2 l/\partial S_i \partial p)^2}{-\partial^2 l/\partial S_i^2} \end{bmatrix},$$

Thus, the matrix P corresponds to

$$P = \begin{bmatrix} V(\hat{\mu}_H) & Cov(\hat{\mu}_H, \hat{p}) \\ Cov(\hat{\mu}_H, \hat{p}) & V(\hat{p}) \end{bmatrix}.$$

Similarly, we derive the matrix Q from Equation (A.12) as follows:

$$Q = \begin{bmatrix} K & \dots & L \\ M & \dots & N \end{bmatrix},$$

in which

$$\begin{aligned}
K &= \frac{(-\partial^2 l/\partial S_1 \partial \mu_H)V(\hat{\mu}_H) + (-\partial^2 l/\partial S_1 \partial p)Cov(\hat{\mu}_H, \hat{p})}{-\partial^2 l/\partial S_1^2} \\
L &= \frac{(-\partial^2 l/\partial S_n \partial \mu_H)V(\hat{\mu}_H) + (-\partial^2 l/\partial S_n \partial p)Cov(\hat{\mu}_H, \hat{p})}{-\partial^2 l/\partial S_n^2} \\
M &= \frac{(-\partial^2 l/\partial S_1 \partial \mu_H)Cov(\hat{\mu}_H, \hat{p}) + (-\partial^2 l/\partial S_1 \partial p)V(\hat{\mu}_H)}{-\partial^2 l/\partial S_1^2} \\
N &= \frac{(-\partial^2 l/\partial S_n \partial \mu_H)Cov(\hat{\mu}_H, \hat{p}) + (-\partial^2 l/\partial S_n \partial p)V(\hat{\mu}_H)}{-\partial^2 l/\partial S_n^2}.
\end{aligned}$$

Finally, the matrix R is calculated from the Equation (A.13) by:

$$R = \begin{bmatrix} A & \dots & B \\ C & \dots & D \\ \dots & \dots & \dots \\ E & \dots & F \end{bmatrix},$$

where

$$\begin{aligned}
A &= 1 - \frac{(-\partial^2 l/\partial S_1 \partial \mu_H)Cov(\hat{S}_1, \hat{\mu}_H) + (-\partial^2 l/\partial S_1 \partial p)Cov(\hat{S}_1, \hat{p})}{-\partial^2 l/\partial S_1^2} \\
B &= -\frac{(-\partial^2 l/\partial S_1 \partial \mu_H)Cov(\hat{S}_n, \hat{\mu}_H) + (-\partial^2 l/\partial S_1 \partial p)Cov(\hat{S}_n, \hat{p})}{-\partial^2 l/\partial S_1^2} \\
C &= -\frac{(-\partial^2 l/\partial S_2 \partial \mu_H)Cov(\hat{S}_1, \hat{\mu}_H) + (-\partial^2 l/\partial S_2 \partial p)Cov(\hat{S}_1, \hat{p})}{-\partial^2 l/\partial S_2^2} \\
D &= -\frac{(-\partial^2 l/\partial S_2 \partial \mu_H)Cov(\hat{S}_n, \hat{\mu}_H) + (-\partial^2 l/\partial S_2 \partial p)Cov(\hat{S}_n, \hat{p})}{-\partial^2 l/\partial S_2^2} \\
E &= -\frac{(-\partial^2 l/\partial S_n \partial \mu_H)Cov(\hat{S}_1, \hat{\mu}_H) + (-\partial^2 l/\partial S_n \partial p)Cov(\hat{S}_1, \hat{p})}{-\partial^2 l/\partial S_n^2} \\
F &= 1 - \frac{(-\partial^2 l/\partial S_n \partial \mu_H)Cov(\hat{S}_n, \hat{\mu}_H) + (-\partial^2 l/\partial S_n \partial p)Cov(\hat{S}_n, \hat{p})}{-\partial^2 l/\partial S_n^2}.
\end{aligned}$$

Then, by making use of the above matrices we derive the following variance and covariance terms:

$$
\begin{aligned}
V(\hat{\mu}_H) &= \frac{1}{C_0}\left[\frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} S_i^2 \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right. \\
&- \left.\sum_{i=1}^{n} \frac{-\frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right]
\end{aligned}
$$

$$
\begin{aligned}
V(\hat{p}) &= \frac{1}{C_0}\left[\frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right]\right. \\
&- \left.\sum_{i=!}^{n} \frac{(\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2})^2}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right]
\end{aligned}
$$

$$
\begin{aligned}
Cov(\hat{\mu}_H, \hat{p}) &= \left[\frac{2v}{k\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right. \\
&+ \sum_{i=1}^{n} \frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \\
&\times \left(-\frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2}\right. \\
&+ \left.\left.\frac{2vp}{k\sigma^2}\sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}\right)\right] \times \frac{1}{C_0},
\end{aligned}
$$

where

$$C_0 = \left(-\frac{\partial^2 l}{\partial \mu_H^2} - \sum_{i=1}^{n} \frac{(\partial^2 l/\partial S_i \partial \mu_H)^2}{-\partial^2 l/\partial S_i^2}\right) \times \left(-\frac{\partial^2 l}{\partial p^2} - \sum_{i=1}^{n} \frac{(\partial^2 l/\partial S_i \partial p)^2}{-\partial^2 l/\partial S_i^2}\right) - \left(-\frac{\partial^2 l}{\partial p \partial \mu_H} - \sum_{i=1}^{n} \frac{(-\partial^2 l/\partial S_i \partial \mu_H)(-\partial^2 l/\partial S_i \partial p)}{-\partial^2 l/\partial S_i^2}\right).$$

Moreover,

$$Cov(\hat{S}_i, \hat{\mu}_H) = \left[\frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \times -V(\hat{\mu}_H)$$

$$- \left[\frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \times Cov(\hat{\mu}_H, \hat{p})$$

and

$$Cov(\hat{S}_i, \hat{p}) = \left[\frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \times -Cov(\hat{\mu}_H, \hat{p})$$

$$- \left[\frac{-\frac{2v}{k\sigma}\sum_{j=1}^{m}\frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m}S_i\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m}\frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}\right] \times V(\hat{p})$$

Also,

$$
\begin{aligned}
V(\hat{S}_i) &= 1 - \left[\frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{S}_i, \hat{\mu}_H) \\
&- \left[\frac{-\frac{2v}{k\sigma}\sum_{j=1}^{m} \frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{S}_i, \hat{p}).
\end{aligned}
$$

Finally,

$$
\begin{aligned}
Cov(\hat{S}_i, \hat{S}_k) &= - \left[\frac{\frac{2v}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{S}_k, \hat{\mu}_H) \\
&- \left[\frac{-\frac{2v}{k\sigma}\sum_{j=1}^{m} \frac{\frac{(b_{i(j)}-pS_i-\mu_H)}{\sigma}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp}{k\sigma^2}\sum_{j=1}^{m} S_i \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}}{\frac{2v}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(a_{i(j)}-S_i-\mu_H)^2}{k\sigma^2}\right)^2} + \frac{2vp^2}{k\sigma^2}\sum_{j=1}^{m} \frac{1-\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}}{\left(1+\frac{(b_{i(j)}-pS_i-\mu_H)^2}{k\sigma^2}\right)^2}} \right] \\
&\times Cov(\hat{S}_k, \hat{p}).
\end{aligned}
$$

APPENDIX B

DERIVATION of ESTIMATORS of ALTERNATIVE MODEL 1

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we consider the following distributional assumption (long-tailed symmetric LTS distribution) on the log-scale:

$$\begin{aligned} \mathrm{PM}_{ij} = a_{ij} &\sim \mathrm{LTS}(S_i + \mu_H, \sigma_i^2), \\ \mathrm{MM}_{ij} = b_{ij} &\sim \mathrm{LTS}(pS_i + \mu_H, \sigma_i^2). \end{aligned} \tag{B.1}$$

Here S_i is the true signal for gene i, μ_H refers to the constant background intensity, and σ_i presents the gene specific standard deviation. Finally, p indicates the fraction of the true signal in MM.

Accordingly, the associated likelihood function is found via:

$$\begin{aligned} L(S_i, \mu_H, p \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_H)^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-pS_i-\mu_H)^2}{2\sigma_i^2}}, \end{aligned}$$

which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m} \left(\frac{1}{\sigma_i}\right)\left(1 + \frac{z_{a_{ij}}^2}{k}\right)^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m} \left(\frac{1}{\sigma_i}\right)\left(1 + \frac{z_{b_{ij}}^2}{k}\right)^{-v},$$

where the shape parameter $v \geq 2, k = 2v-3, a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$.

$$\begin{aligned} z_{a_{ij}} &= \frac{(a_{ij} - S_i - \mu_H)}{\sigma_i} \\ z_{b_{ij}} &= \frac{(b_{ij} - pS_i - \mu_H)}{\sigma_i}. \end{aligned}$$

Then, the logarithm of L is derived as:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right]+\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i)+ln\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right]+\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i)+ln\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i)+vln\left(1+\frac{z_{a_{ij}}^2}{k}\right)\right]-\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i)+vln\left(1+\frac{z_{b_{ij}}^2}{k}\right)\right].
\end{aligned}$$

In order to get the MLE of model parameters, we take the first derivatives of lnL with respect to each parameter as follows:

$$\begin{aligned}
\frac{\partial\ln L}{\partial\mu_H} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial p} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i(b_{ij}-pS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}}-\frac{2vp}{k}\sum_{j=1}^{m}\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial\sigma_i} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2})}\frac{2v}{k}\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i^3(1+\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2})}\frac{2v}{k}\right],
\end{aligned}$$

where

$$\begin{aligned}
z_{a_{ij}} &= \frac{(a_{ij}-S_i-\mu_H)}{\sigma_i} \\
z_{b_{ij}} &= \frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i}.
\end{aligned}$$

Then, we approximate the common nonlinear functions $g(z)=\frac{z}{1+\frac{z^2}{k}}$ for PM and MM as:

$$g(z_{a_{ij}})=\frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i}}{1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} \text{ and } g(z_{b_{ij}})=\frac{\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i}}{1+\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}},$$

respectively, by the first order Taylor expansions via:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}}$$

and

$$g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

where

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2},$$

when $a_{i(j)}$ and $b_{i(j)}$ represent the ordered PM and MM with respect to the probes j, i.e., the concomitant. Hereby, we can present the partial derivatives of MMLE as below:

$$\frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \tag{B.2}$$

$$\frac{\partial \ln L}{\partial p} = \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \tag{B.3}$$

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2vp}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \tag{B.4}$$

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{b_{i(j)}}). \tag{B.5}$$

Finally, to obtain the estimates of parameters, we set Equations (B.2) - (B.5) to zero. Accordingly, from Equation (B.2):

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) &= 0 \\
\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) &= 0 \\
\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}(\alpha_j + \beta_j z_{a_{i(j)}} + \alpha_j + \beta_j z_{b_{i(j)}}) &= 0 \\
\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} &= 0 \\
\sum_{i=1}^{n}\frac{1}{\sigma_i}\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{i=1}^{n}\frac{1}{\sigma_i}\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} &= 0,
\end{aligned}$$

by taking $\sum_{j=1}^{m} \alpha_j = 0$ due to the symmetry. Then,

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} = 0 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)} - S_i - \mu_H)}{\sigma_i^2} + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(b_{i(j)} - pS_i - \mu_H)}{\sigma_i^2} = 0.
\end{aligned}$$

So

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{a_{i(j)}}{\sigma^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2} - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{\mu_H}{\sigma^2} = \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{b_{i(j)}}{\sigma^2} + p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2}2\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{\mu_H}{\sigma^2} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)} + b_{i(j)})}{\sigma^2} - (p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2}.
\end{aligned}$$

As a result, $\hat{\mu}_H$ is found by:

$$\hat{\mu}_H = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)}+b_{i(j)})}{\sigma^2} - (p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2}}{2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma^2}}. \tag{B.6}$$

On the other hand, from Equation (B.4):

$$\begin{aligned}
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2vp}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) = 0 \\
0 &= \frac{2v}{k}\frac{1}{\sigma_i}\left[\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + pg(z_{b_{i(j)}}))\right] \\
0 &= \frac{2v}{k\sigma_i}\sum_{j=1}^{m}((\alpha_j + \beta_j z_{a_{i(j)}}) + \frac{2v}{k\sigma_i}\sum_{j=1}^{m}((\alpha_j + \beta_j z_{b_{i(j)}}) \\
0 &= \frac{1}{\sigma_i^2}\sum_{j=1}^{m}\beta_j(a_{i(j)} - S_i - \mu_H) + \frac{p}{\sigma_i^2}\sum_{j=1}^{m}\beta_j(b_{i(j)} - pS_i - \mu_H) \\
0 &= \sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j S_i - \mu_H\sum_{j=1}^{m}\beta_j + p\sum_{j=1}^{m}\beta_j b_{i(j)} \\
&- p^2\sum_{j=1}^{m}\beta_j S_i - p\mu_H\sum_{j=1}^{m}\beta_j \\
&= (p^2+1)S_i\sum_{j=1}^{m}\beta_j = \sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_H\sum_{j=1}^{m}\beta_j.
\end{aligned}$$

Hence, we get the estimate of the true signal for each gene i via:

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_H \sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j}. \tag{B.7}$$

But by substituting B.7 into B.6, we can also define $\hat{\mu}_H$ as follows:

$$\begin{aligned}\hat{\mu}_H &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)}+b_{i(j)})}{\sigma^2}}{2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma^2}} \\ &- \frac{(p+1)\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}+p\sum_{j=1}^{m}\beta_j b_{i(j)}-(p+1)\mu_H\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j}}{2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma^2}}.\end{aligned}$$

So,

$$\hat{\mu}_H = \frac{p\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{(ij)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{(ij)}}{(p^2+1)\sum_{j=1}^{m}\beta_j}. \tag{B.8}$$

Similarly, by substituting Equation B.8 into Equation B.7, an alternative form of $\hat{S}_i$ can be written as:

$$\begin{aligned}\hat{S}_i &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p^2+1)\sum_{j=1}^{m}\beta_j} - \frac{(p+1)\left[\frac{p\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{(ij)}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{(ij)}}{(p^2+1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j} \\ &= \frac{1}{(p^2+1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j} + \frac{p}{(p^2+1)}\frac{\sum_{j=1}^{m}\beta_j b_{i(j)}}{\sum_{j=1}^{m}\beta_j} \\ &- \frac{p(p+1)}{(p-1)(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{(ij)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}} + \frac{(p+1)}{(p-1)(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{(ij)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}}\end{aligned}$$

Accordingly,

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p^2+1)\sum_{j=1}^{m}\beta_j} - \frac{\frac{\frac{p(p+1)}{(p-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{i(j)}-\frac{(p+1)}{(p-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}}{(p^2+1)\sum_{j=1}^{m}\beta_j}. \tag{B.9}$$

On the other side, from Equation (B.3), we can get the estimate of common fraction p as below:

$$
\begin{aligned}
\frac{\partial \ln L}{\partial p} = \frac{2v}{k} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{S_i}{\sigma_i} g(z_{b_{i(j)}}) &= 0 \\
\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{S_i}{\sigma_i} g(z_{b_{i(j)}}) &= 0 \\
\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{S_i}{\sigma_i} \alpha_j + \sum_{j=1}^{m} \beta_j \frac{(b_{i(j)} - pS_i - \mu_H)}{\sigma_i} \frac{S_i}{\sigma_i} &= 0 \\
\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i}{\sigma_i^2} b_{i(j)} - p \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i^2}{\sigma_i^2} - \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_H \frac{\beta_j S_i}{\sigma_i^2}. &= 0
\end{aligned}
$$

Hereby,

$$
\hat{p} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i}{\sigma_i^2} b_{i(j)} - \sum_{i=1}^{n} \sum_{j=1}^{m} \mu_H \frac{\beta_j S_i}{\sigma_i^2}}{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i^2}{\sigma_i^2}}.
$$

We can write the above equation as follows, too:

$$
\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i}{\sigma_i^2} b_{i(j)} - \hat{p} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i^2}{\sigma_i^2} - \mu_H \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j S_i}{\sigma_i^2} = 0. \tag{B.10}
$$

Then, by substituting Equations (B.8) and (B.9) into Equation (B.10):

$$
\begin{aligned}
& \left[\frac{\sum_{j=1}^{m} \beta_j a_{i(j)} + p \sum_{j=1}^{m} \beta_j b_{i(j)} - \frac{\frac{p(p+1)}{(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} a_{i(j)} - \frac{(p+1)}{(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} b_{i(j)}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}{(p^2+1) \sum_{j=1}^{m} \beta_j} \right] \\
\times & \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} b_{i(j)} \\
- & \left[\frac{\sum_{j=1}^{m} \beta_j a_{i(j)} + p \sum_{j=1}^{m} \beta_j b_{i(j)} - \frac{\frac{p(p+1)}{(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} a_{i(j)} - \frac{(p+1)}{(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} b_{i(j)}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}{(p^2+1) \sum_{j=1}^{m} \beta_j} \right]^2 \\
\times & p \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} \\
- & \left[\frac{p \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} a_{(ij)} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} b_{(ij)}}{(p^2+1) \sum_{j=1}^{m} \beta_j} \right] \\
\times & \left[\frac{\sum_{j=1}^{m} \beta_j a_{i(j)} + p \sum_{j=1}^{m} \beta_j b_{i(j)} - \frac{\frac{p(p+1)}{(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} a_{i(j)} - \frac{(p+1)}{(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} b_{i(j)}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}{(p^2+1) \sum_{j=1}^{m} \beta_j} \right] \\
\times & \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i^2} = 0.
\end{aligned}
$$

Finally, from Equation (B.5), we can obtain MLE of σ for each gene i by setting this expression to zero via:

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \sigma_i} &= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} g(z_{b_{i(j)}}) \\
&= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} (g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) = 0
\end{aligned}
$$

Also,

$$\begin{aligned}
\frac{2m}{\sigma_i} &= \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\sigma_i &= \frac{v}{km}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\frac{km\sigma_i}{v} &= \sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \sum_{j=1}^{m}\left(\alpha_j + \beta_j\frac{a_{i(j)} - S_i - \mu_H}{\sigma_i} + \alpha_j + \beta_j\frac{b_{i(j)} - pS_i - \mu_H}{\sigma_i}\right) \\
\frac{km\sigma_i^2}{v} &= \sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j S_i - \sum_{j=1}^{m}\beta_j\mu_H + \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j pS_i - \sum_{j=1}^{m}\beta_j\mu_H \\
&= \sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)}) - (p+1)S_i\sum_{j=1}^{m}\beta_j - 2\mu_H\sum_{j=1}^{m}\beta_j.
\end{aligned}$$

Hence,

$$\hat{\sigma_i}^2 = \frac{km}{v}\left(\sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)}) - (p+1)S_i\sum_{j=1}^{m}\beta_j - 2\mu_H\sum_{j=1}^{m}\beta_j\right).$$

Finally, by substituting Equations (B.8) and (B.9) into the above equation, we get the most simple form of σ_i, which is the non-linear functions of p, resulting in no explicit expressions for the model parameter of Equation (B.1):

$$
\begin{aligned}
\hat{\sigma_i}^2 &= \frac{km}{v}\left[\sum_{j=1}^{m}\beta_j(a_{i(j)}+b_{i(j)})\right] \\
&- \frac{km(p+1)}{v}\left[\frac{1}{(p^2+1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j}+\frac{p}{(p^2+1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j}\right] \\
&- \frac{km(p+1)}{v}\left[\frac{p(p+1)}{(p-1)(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{i(j)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}}\right. \\
&+ \left.\frac{(p+1)}{(p-1)(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}}\right] \\
&- \frac{2km}{v}\sum_{j=1}^{m}\beta_j\left[\frac{p}{(p-1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{i(j)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}}+\frac{1}{(p-1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}}\right] \\
\hat{\sigma_i}^2 &= \frac{km}{v}\left[\frac{p(p-1)}{(p^2+1)}\sum_{j=1}^{m}\beta_j a_{i(j)}-\frac{(p-1)}{(p^2+1)}\sum_{j=1}^{m}\beta_j b_{i(j)}\right. \\
&- \left.\frac{p(p-1)}{(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}a_{i(j)}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\right]+\frac{km}{v}\left[\frac{(p-1)}{(p^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}\right].
\end{aligned}
$$

APPENDIX C

DERIVATION of ESTIMATORS of ALTERNATIVE MODEL 2

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, the following distributional assumption under the long-tailed symmetric(LTS) density is considered on the log-scale:

$$\mathrm{PM}_{ij} = a_{ij} \sim \mathrm{LTS}(S_i + \mu_H, \sigma_i^2)$$

and

$$\mathrm{MM}_{ij} = b_{ij} \sim \mathrm{LTS}(p_i S_i + \mu_H, \sigma_i^2),$$

where S_i and μ_H are the gene specific true signal and background intensity, respectively, and σ_i denotes the standard deviation for gene i as used in previous alternative models. On the other hand, p_i presents the fraction of the true signal in MM for each gene i.

Thereby, the corresponding likelihood is found via:

$$\begin{aligned} L(S_i, \mu_H, p_i \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_H)^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-p_i S_i-\mu_H)^2}{2\sigma_i^2}}, \end{aligned}$$

which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m} (\frac{1}{\sigma_i})(1 + \frac{z_{a_{ij}}^2}{k})^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m} (\frac{1}{\sigma_i})(1 + \frac{z_{b_{ij}}^2}{k})^{-v},$$

where the shape parameter $v \geq 2$, $k = 2v - 3$, degree of freedom $d = 2v - 1$, and finally a and b refer to nm-dimensional vectors $a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$, in order.

$$\begin{aligned} z_{a_{ij}} &= \frac{(a_{ij} - S_i - \mu_H)}{\sigma_i}, \\ z_{b_{ij}} &= \frac{(b_{ij} - p_i S_i - \mu_H)}{\sigma_i}. \end{aligned}$$

Then, the logarithm of L is found as:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n}\sum_{j=1}^{m} \ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right] + \sum_{i=1}^{n}\sum_{j=1}^{m} \ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i) + ln\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i) + ln\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i) + vln\left(1+\frac{z_{a_{ij}}^2}{k}\right)\right] \\
&- \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i) + vln\left(1+\frac{z_{b_{ij}}^2}{k}\right)\right].
\end{aligned}$$

To obtain the MLE of model parameters, we take the first derivatives of $\ln L$ with respect to each parameter as below:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_H} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} + \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(b_{ij}-p_iS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_H)^2}{k\sigma_i^2}} \\
\frac{\partial \ln L}{\partial p_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{S_i(b_{ij}-p_iS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_H)^2}{k\sigma_i^2}} \\
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} + \frac{2v}{k}\sum_{j=1}^{m}\frac{-p_i(b_{ij}-p_iS_i-\mu_H)}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_H)^2}{k\sigma_i^2}} \\
\frac{\partial \ln L}{\partial \sigma_i} &= -\sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2})}\frac{2v}{k}\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-p_iS_i-\mu_H)}{\sigma_i^3(1+\frac{(b_{ij}-p_iS_i-\mu_H)^2}{k\sigma_i^2})}\frac{2v}{k}\right],
\end{aligned}$$

where

$$z_{a_{ij}} = \frac{(a_{ij}-S_i-\mu_H)}{\sigma_i}$$

and

$$z_{b_{ij}} = \frac{(b_{ij}-p_iS_i-\mu_H)}{\sigma_i}.$$

Then, we approximate the common nonlinear functions $g(z) = \frac{z}{1+\frac{z^2}{k}}$ for PM and MM as:

$$g(z_{a_{ij}}) = \frac{\frac{(a_{ij}-S_i-\mu_H)}{\sigma_i}}{1+\frac{(a_{ij}-S_i-\mu_H)^2}{k\sigma_i^2}} \text{ and } g(z_{b_{ij}}) = \frac{\frac{(b_{ij}-pS_i-\mu_H)}{\sigma_i}}{1+\frac{(b_{ij}-p_iS_i-\mu_H)^2}{k\sigma_i^2}},$$

respectively, by the first order Taylor expansions via:

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

where

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2},$$

for the standardized and ordered PM and MM intensities (in increasing magnitude), in order, with respect to the probes in each gene i. Moreover, here $t_(j)$ refers to the ordered associate student-t quantile for each probe j ($j = 1, \ldots, m$).

Then, the partial derivatives of MLE can be shown as follows:

$$\frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \quad \text{(C.1)}$$

$$\frac{\partial \ln L}{\partial p} = \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \quad \text{(C.2)}$$

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{j=1}^{m}\frac{p_i}{\sigma_i}g(z_{b_{i(j)}}) \quad \text{(C.3)}$$

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{b_{i(j)}}). \quad \text{(C.4)}$$

Finally, in order to get the estimates of parameters, we set Equations (C.1) - (C.4) to zero. Thereby, from Equation (C.1) we derive:

$$\frac{\partial \ln L}{\partial \mu_H} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{\sigma_i}(\alpha_j + \beta_j z_{a_{i(j)}} + \alpha_j + \beta_j z_{b_{i(j)}}) = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} = 0$$

$$\sum_{i=1}^{n}\frac{1}{\sigma_i}\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{i=1}^{n}\frac{1}{\sigma_i}\sum_{j=1}^{m}\alpha_j + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} = 0,$$

by taking $\sum_{j=1}^{m}\alpha_j = 0$ due to the symmetry. Then,

$$\frac{\partial \ln L}{\partial \mu_H} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)} - S_i - \mu_H)}{\sigma_i^2} + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(b_{i(j)} - p_i S_i - \mu_H)}{\sigma_i^2} = 0$$

Accordingly,

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{a_{i(j)}}{\sigma^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2} - 2\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{\mu_H}{\sigma^2} =$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{b_{i(j)}}{\sigma^2} + p\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2}.$$

So,

$$2\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{\mu_H}{\sigma^2} = \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)} + b_{i(j)})}{\sigma^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2}(p_i + 1).$$

Finally, $\hat{\mu}_H$ can be derived as:

$$\hat{\mu}_H = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)}+b_{i(j)})}{\sigma^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{S_i}{\sigma^2}(p_i + 1)}{2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma^2}}. \quad \text{(C.5)}$$

On the other side, by setting Equation (C.3) to zero

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2vp}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}).$$

Thereby,

$$\begin{aligned}
\frac{2v}{k}\frac{1}{\sigma_i}\left[\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + pg(z_{b_{i(j)}}))\right] &= 0 \\
\frac{2v}{k\sigma_i}\sum_{j=1}^{m}((\alpha_j + \beta_j z_{a_{i(j)}}) + \frac{2vp_i}{k\sigma_i}\sum_{j=1}^{m}((\alpha_j + \beta_j z_{b_{i(j)}}) &= 0 \\
\frac{1}{\sigma_i^2}\sum_{j=1}^{m}\beta_j(a_{i(j)} - S_i - \mu_H) + \frac{p_i}{\sigma_i^2}\sum_{j=1}^{m}\beta_j(b_{i(j)} - p_i S_i - \mu_H) &= 0 \\
\sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j S_i - \mu_H\sum_{j=1}^{m}\beta_j + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - p_i^2\sum_{j=1}^{m}\beta_j S_i - p_i\mu_H\sum_{j=1}^{m}\beta_j &= 0.
\end{aligned}$$

So,

$$(p^2+1)S_i\sum_{j=1}^{m}\beta_j = \sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - (p_i+1)\mu_H\sum_{j=1}^{m}\beta_j,$$

we get the estimate of the true signal for each gene i via:

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - (p_i+1)\mu_H\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}. \tag{C.6}$$

However, by substituting (C.6) into (C.5), we can also define $\hat{\mu}_H$ in a different way as below:

$$\hat{\mu}_H = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j\frac{(a_{i(j)}+b_{i(j)})}{\sigma_i}}{2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(p_i+1)\frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - (p_i+1)\mu_H\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}}{2\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}} \tag{C.7}$$

$$\begin{aligned}
\hat{\mu}_H &= \frac{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\sum_{i=1}^{n}\beta_j} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j a_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}} \\
&- \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)p_i}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j b_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}.
\end{aligned} \tag{C.8}$$

Likewise, by substituting (C.7) into (C.6), another form of $\hat{S}_i$ can be written via:

$$\begin{aligned}
\hat{S}_i &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - (p_i+1)\mu_H\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j} \\
&= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p_i^2+1)\sum_{j=1}^{m}\beta_j} \\
&- \frac{(p_i+1)\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\sum_{i=1}^{n}\beta_j} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j a_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}\right]\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j} \\
&+ \frac{(p_i+1)\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)p_i}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j b_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}\right]\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}.
\end{aligned}$$

Hereby,

$$\begin{aligned}
\hat{S}_i &= \frac{1}{(p_i^2+1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j} + \frac{p_i}{(p_i^2+1)}\frac{\sum_{j=1}^{m}\beta_j b_{i(j)}}{\sum_{j=1}^{m}\beta_j} \\
&- \frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\left(2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)\sum_{j=1}^{m}\beta_j} \\
&- \frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_i+1}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}a_{i(j)}}{\left(2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)} \\
&- \frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_i(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}b_{i(j)}}{\left(2\sum_{i=1}^{n}\frac{1}{\sigma_i} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)}.
\end{aligned} \tag{C.9}$$

On the other side, from Equation (C.2), the estimator of the common fraction p for each gene i, can be derives as follows:

$$\frac{\partial \ln L}{\partial p_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) = \frac{S_i}{\sigma_i}\sum_{j=1}^{m}g(z_{b_{i(j)}}).$$

As a result,

$$\begin{aligned}
\sum_{j=1}^{m} g(z_{b_{i(j)}}) = \sum_{j=1}^{m} \alpha_j + \sum_{j=1}^{m} \beta_j \frac{(b_{i(j)} - p_i S_i - \mu_H)}{\sigma_i} &= 0 \\
\sum_{j=1}^{m} \beta_j \frac{(b_{i(j)} - p_i S_i - \mu_H)}{\sigma_i} = \sum_{j=1}^{m} \beta_j (b_{i(j)} - p_i S_i - \mu_H) &= 0 \\
\sum_{j=1}^{m} \beta_j b_{i(j)} - \sum_{j=1}^{m} \beta_j p_i S_i - \sum_{j=1}^{m} \beta_j \mu_H) &= 0.
\end{aligned}$$

Thereby,

$$\hat{p} = \frac{\sum_{j=1}^{m} b_{i(j)} - \mu_H \sum_{j=1}^{m} \beta_j}{S_i \sum_{j=1}^{m} \beta_j}.$$

We can write the above equation as follows, too:

$$\begin{aligned}
\hat{p} S_i \sum_{j=1}^{m} \beta_j - \sum_{j=1}^{m} b_{i(j)} + \mu_H \sum_{j=1}^{m} \beta_j = 0 \\
\hat{p} S_i - \frac{\sum_{j=1}^{m} b_{i(j)}}{\sum_{j=1}^{m} \beta_j} + \mu_H = 0. \qquad \text{(C.10)}
\end{aligned}$$

(C.11)

Then, by substituting Equations (C.7) and (C.9) into Equation (C.10):

$$\begin{aligned}
& \hat{p}_i \left[\frac{1}{(p_i^2+1)} \frac{\sum_{j=1}^{m} \beta_j a_{i(j)}}{\sum_{j=1}^{m} \beta_j} + \frac{p_i}{(p_i^2+1)} \frac{\sum_{j=1}^{m} \beta_j b_{i(j)}}{\sum_{j=1}^{m} \beta_j} \right] \\
- \quad & \hat{p}_i \left[\frac{p_i+1}{(p_i^2+1)} \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i} (a_{i(j)} + b_{i(j)})}{(2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}) \sum_{j=1}^{m} \beta_j} \right] \\
- \quad & \hat{p}_i \left[\frac{p_i+1}{(p_i^2+1)} \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_i+1}{(p_i^2+1)} \frac{\beta_j}{\sigma_i} a_{i(j)}}{(2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})} \right] \\
- \quad & \hat{p}_i \left[\frac{p_i+1}{(p_i^2+1)} \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_i(p_i+1)}{(p_i^2+1)} \frac{\beta_j}{\sigma_i} b_{i(j)}}{(2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})} \right] - \frac{\sum_{j=1}^{m} b_{i(j)}}{\sum_{j=1}^{m} \beta_j} \\
+ \quad & \frac{\frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i} (a_{i(j)} + b_{i(j)})}{\sum_{i=1}^{n} \beta_j} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)}{(p_i^2+1)} \frac{\beta_j}{\sigma_i} \sum_{i=1}^{n} \beta_j a_{i(j)}}{2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}} \\
- \quad & \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)p_i}{(p_i^2+1)} \frac{\beta_j}{\sigma_i} \sum_{i=1}^{n} \beta_j b_{i(j)}}{2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}} = 0.
\end{aligned}$$

By solving the above equation we get

$$
\begin{aligned}
p_i \frac{\sum_{j=1}^{m} \beta_j a_{i(j)}}{\sum_{j=1}^{m} \beta_j} &- \sum_{j=1}^{m} \beta_j b_{i(j)} - (p_i - 1) \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j}{\sigma_i} (a_{i(j)} + b_{i(j)})}{\left(2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right) \sum_{j=1}^{m} \beta_j} \\
&- (2p_i^2 + p_i + 1) \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_i+1}{(p_i^2+1)} \frac{\beta_j}{\sigma_i} a_{i(j)}}{\left(2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)} \\
&- (2p_i^2 + p_i + 1) \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_i(p_i+1)}{(p_i^2+1)} \frac{\beta_j}{\sigma_i} b_{i(j)}}{\left(2 \sum_{i=1}^{n} \frac{1}{\sigma_i} - \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}\right)} = 0.
\end{aligned}
$$

Finally, from Equation (C.4), we can obtain MLE of σ for each gene i by setting this expression to zero via:

$$
\begin{aligned}
\frac{\partial \ln L}{\partial \sigma_i} &= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} g(z_{b_{i(j)}}) = 0 \\
&= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} (g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\frac{2m}{\sigma_i} &= \frac{2v}{k\sigma_i^2} \sum_{j=1}^{m} (g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\sigma_i &= \frac{v}{km} \sum_{j=1}^{m} (g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\frac{km\sigma_i}{v} &= \sum_{j=1}^{m} (g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \sum_{j=1}^{m} \left(\alpha_j + \beta_j \frac{a_{i(j)} - S_i - \mu_H}{\sigma_i} + \alpha_j + \beta_j \frac{b_{i(j)} - p_i S_i - \mu_H}{\sigma_i}\right) \\
\frac{km\sigma_i^2}{v} &= \sum_{j=1}^{m} \beta_j a_{i(j)} - \sum_{j=1}^{m} \beta_j S_i - \sum_{j=1}^{m} \beta_j \mu_H + \sum_{j=1}^{m} \beta_j b_{i(j)} - \sum_{j=1}^{m} \beta_j p_i S_i - \sum_{j=1}^{m} \beta_j \mu_H \\
&= \sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p_i + 1) S_i \sum_{j=1}^{m} \beta_j - 2\mu_H \sum_{j=1}^{m} \beta_j.
\end{aligned}
$$

Hence,

$$
\hat{\sigma_i}^2 = \frac{km}{v} \sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p_i + 1) S_i \sum_{j=1}^{m} \beta_j - 2\mu_H \sum_{j=1}^{m} \beta_j.
$$

Finally, by substituting Equations (C.7) and (C.9) into the above equation, we get the most simple form of σ_i, which is the non-linear functions of p_i, leading to no explicit solutions for the model parameter of Equation (C.1):

$$
\begin{aligned}
\hat{\sigma_i}^2 &= \frac{km}{v}\sum_{j=1}^{m}\beta_j(a_{i(j)}+b_{i(j)}) \\
&- \frac{(p_i+1)v}{km}\left[\frac{1}{(p_i^2+1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j}+\frac{p_i}{(p_i^2+1)}\frac{\sum_{j=1}^{m}\beta_j b_{i(j)}}{\sum_{j=1}^{m}\beta_j}\right] \\
&+ \frac{(p_i+1)v}{km}\left[\frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{(2\sum_{i=1}^{n}\frac{1}{\sigma_i}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})\sum_{j=1}^{m}\beta_j}\right] \\
&+ \frac{(p_i+1)v}{km}\left[\frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_i+1}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}a_{i(j)}}{(2\sum_{i=1}^{n}\frac{1}{\sigma_i}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})}\right] \\
&+ \frac{(p_i+1)v}{km}\left[\frac{p_i+1}{(p_i^2+1)}\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_i(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}b_{i(j)}}{(2\sum_{i=1}^{n}\frac{1}{\sigma_i}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i})}\right] \\
&- 2\sum_{j=1}^{m}\beta_j\left[\frac{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i}(a_{i(j)}+b_{i(j)})}{\sum_{i=1}^{n}\beta_j}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j a_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}\right] \\
&+ 2\sum_{j=1}^{m}\beta_j\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)p_i}{(p_i^2+1)}\frac{\beta_j}{\sigma_i}\sum_{i=1}^{n}\beta_j b_{i(j)}}{2\sum_{i=1}^{n}\frac{1}{\sigma_i}-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(p_i+1)^2}{(p_i^2+1)\sigma_i}}\right]=0.
\end{aligned}
$$

From this final expression, it is seen that similar to the first alternative model, Model 2 also has none explicit expression for the model estimators.

APPENDIX D

DERIVATION of ESTIMATORS of ALTERNATIVE MODEL 3

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we consider the following distributional assumption under the long tailed symmetric (LTS) on the log-scale:

$$\begin{aligned} \text{PM}_{ij} &= a_{ij} \sim \text{LTS}(S_i + \mu_{Hi}, \sigma_i^2) \\ \text{MM}_{ij} &= b_{ij} \sim \text{LTS}(pS_i + \mu_{Hi}, \sigma_i^2), \end{aligned} \tag{D.1}$$

in which S_i, p and μ_H describe the true signal in gene i, constant fraction of true signal in MM, and gene-specific background intensities, respectively, as described beforehand. Finally, σ_i^2 is the gene-specific variance component.

Hereby, the associated likelihood is found via:

$$\begin{aligned} L(S_i, \mu_{Hi}, p \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_{Hi})^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{2\sigma_i^2}}, \end{aligned}$$

for $a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$, and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$, which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m} \left(\frac{1}{\sigma_i}\right)\left(1 + \frac{z_{a_{ij}}^2}{k}\right)^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m} \left(\frac{1}{\sigma_i}\right)\left(1 + \frac{z_{b_{ij}}^2}{k}\right)^{-v}$$

and under the shape parameter $v \geq 2$ and $k = 2v - 3$, as well as:

$$z_{a_{ij}} = \frac{(a_{ij} - S_i - \mu_{Hi})}{\sigma_i} \text{ and } z_{b_{ij}} = \frac{(b_{ij} - pS_i - \mu_{Hi})}{\sigma_i}.$$

Then, the logarithm of L is derived as:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right]+\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i)+\ln\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right]+\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i)+\ln\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i)+v\ln\left(1+\frac{z_{a_{ij}}^2}{k}\right)\right]-\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i)+v\ln\left(1+\frac{z_{b_{ij}}^2}{k}\right)\right].
\end{aligned}$$

In order to get the MLE of model parameters, we take the first derivatives of $\ln L$ with respect to each parameter as follows:

$$\begin{aligned}
\frac{\partial\ln L}{\partial\mu_{Hi}} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{j=1}^{m}\frac{(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial p} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{j=1}^{m}\frac{-p(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial\sigma_i} &= \sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right] \\
&+ \sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right],
\end{aligned}$$

where

$$\begin{aligned}
z_{a_{ij}} &= \frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i} \\
z_{b_{ij}} &= \frac{(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i}.
\end{aligned}$$

Then, by approximating the common nonlinear functions $g(z)=\frac{z}{1+\frac{z^2}{k}}$ for PM and MM via:

$$g(z_{a_{ij}})=\frac{\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i}}{1+\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}$$

and

$$g(z_{b_{ij}})=\frac{\frac{(b_{ij}-pS_i-\mu_{Hi})}{\sigma_i}}{1+\frac{(b_{ij}-pS_i-\mu_{Hi})^2}{k\sigma_i^2}},$$

by the first order Taylor expansions we get

$$g(z_{a_{i(j)}}) = \alpha_j + \beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}}) = \alpha_j + \beta_j z_{b_{i(j)}},$$

where

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2},$$

under the probe based ordered values of z_{ij} for each gene i. Here $t_{(i)}$ indicates the quantile of the student-t density for the jth probe, as used other alternative models. We can express the partial derivatives of MMLE as below:

$$\frac{\partial \ln L}{\partial \mu_{Hi}} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \quad \text{(D.2)}$$

$$\frac{\partial \ln L}{\partial p} = \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \quad \text{(D.3)}$$

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2vp}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \quad \text{(D.4)}$$

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{b_{i(j)}}). \quad \text{(D.5)}$$

Finally, to obtain the estimates of parameters, we set Equations (D.2 - D.5) to zero. Thereby, from Equation (D.2) we get

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_{Hi}} = \sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) &= 0 \\
\sum_{j=1}^{m}\frac{1}{\sigma_i}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) &= 0 \\
\sum_{j=1}^{m}\frac{1}{\sigma_i}(\alpha_j + \beta_j z_{a_{i(j)}} + \alpha_j + \beta_j z_{b_{i(j)}}) &= 0 \\
\sum_{j=1}^{m}\frac{\alpha_j}{\sigma_i} + \sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{j=1}^{m}\frac{\alpha_j}{\sigma_i} + \sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} &= 0 \\
\frac{1}{\sigma_i}\sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \frac{1}{\sigma_i}\sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\frac{\beta_j z_{b_{i(j)}}}{\sigma_i} &= 0,
\end{aligned}$$

by taking $\sum_{j=1}^{m}\alpha_j = 0$ because of the symmetry. Then,

$$\begin{aligned}\frac{\partial \ln L}{\partial \mu_{Hi}} &= \sum_{j=1}^{m} \frac{\beta_j z_{a_{i(j)}}}{\sigma_i} + \sum_{j=1}^{m} \frac{\beta_j z_{b_{i(j)}}}{\sigma_i} = 0 \\ &= \sum_{j=1}^{m} \beta_j \frac{(a_{i(j)} - S_i - \mu_{Hi})}{\sigma_i^2} + \sum_{j=1}^{m} \beta_j \frac{(b_{i(j)} - pS_i - \mu_{Hi})}{\sigma_i^2}.\end{aligned}$$

Accordingly,

$$\begin{aligned}\sum_{j=1}^{m} \beta_j \frac{a_{i(j)}}{\sigma^2} - \sum_{j=1}^{m} \beta_j \frac{S_i}{\sigma^2} - \sum_{j=1}^{m} \beta_j \frac{\mu_{Hi}}{\sigma^2} &= -\sum_{j=1}^{m} \beta_j \frac{b_{i(j)}}{\sigma^2} + p \sum_{j=1}^{m} \beta_j \frac{S_i}{\sigma^2} + \sum_{j=1}^{m} \beta_j \frac{\mu_{Hi}}{\sigma^2} \\ 2 \sum_{j=1}^{m} \beta_j \frac{\mu_{Hi}}{\sigma^2} &= \sum_{j=1}^{m} \beta_j \frac{(a_{i(j)} + b_{i(j)})}{\sigma^2} - (p+1) \sum_{j=1}^{m} \beta_j \frac{S_i}{\sigma^2}.\end{aligned}$$

Finally, $\hat{\mu}_{Hi}$ is found as:

$$\hat{\mu}_{Hi} = \frac{\sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p+1) S_i \sum_{j=1}^{m} \beta_j}{2 \sum_{j=1}^{m} \beta_j}. \tag{D.6}$$

On the other hand, from Equation (D.4):

$$\begin{aligned}\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k} \sum_{j=1}^{m} \frac{1}{\sigma_i} g(z_{a_{i(j)}}) + \frac{2vp}{k} \sum_{j=1}^{m} \frac{1}{\sigma_i} g(z_{b_{i(j)}}) = 0 \\ &= \frac{2v}{k} \frac{1}{\sigma_i} \left[\sum_{j=1}^{m} (g(z_{a_{i(j)}}) + p g(z_{b_{i(j)}})) \right] \\ &= \frac{2v}{k\sigma_i} \sum_{j=1}^{m} ((\alpha_j + \beta_j z_{a_{i(j)}}) + \frac{2v}{k\sigma_i} \sum_{j=1}^{m} ((\alpha_j + \beta_j z_{b_{i(j)}}) \\ &= \frac{1}{\sigma_i^2} \sum_{j=1}^{m} \beta_j (a_{i(j)} - S_i - \mu_{Hi}) + \frac{p}{\sigma_i^2} \sum_{j=1}^{m} \beta_j (b_{i(j)} - pS_i - \mu_{Hi}) \\ &= \sum_{j=1}^{m} \beta_j a_{i(j)} - \sum_{j=1}^{m} \beta_j S_i - \mu_{Hi} \sum_{j=1}^{m} \beta_j + p \sum_{j=1}^{m} \beta_j b_{i(j)} \\ &- p^2 \sum_{j=1}^{m} \beta_j S_i - p \mu_{Hi} \sum_{j=1}^{m} \beta_j.\end{aligned}$$

So,

$$(p^2+1) S_i \sum_{j=1}^{m} \beta_j = \sum_{j=1}^{m} \beta_j a_{i(j)} + p \sum_{j=1}^{m} \beta_j b_{i(j)} - (p+1) \mu_{Hi} \sum_{j=1}^{m} \beta_j.$$

Then, the estimate of the true signal for each gene i is derived as below:

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_{Hi}\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j}. \quad \text{(D.7)}$$

But by substituting D.7 into D.6, we can also define $\hat{\mu}_H$ as follows:

$$\hat{\mu}_{Hi} = \frac{\sum_{j=1}^{m}\beta_j(a_{i(j)}+b_{i(j)})}{2\sum_{j=1}^{m}\beta_j} - \frac{(p+1)\left[\frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_{Hi}\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j}{2\sum_{j=1}^{m}\beta_j}$$

Therefore,

$$\hat{\mu}_{Hi} = \frac{p\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{j=1}^{m}\beta_j b_{(ij)}}{(p^2+1)\sum_{j=1}^{m}\beta_j}. \quad \text{(D.8)}$$

Similarly, by substituting Equation D.8 into Equation D.7, an alternative form of $\hat{S}_i$ can be written as:

$$\begin{aligned}
\hat{S}_i &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)} - (p+1)\mu_{Hi}\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j} \\
&= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p^2+1)\sum_{j=1}^{m}\beta_j} \\
&- \frac{(p+1)\left[\frac{p\sum_{j=1}^{m}\beta_j(a_{(ij)}-\sum_{j=1}^{m}\beta_j(b_{(ij)})}{(p^2+1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j}{(p^2+1)\sum_{j=1}^{m}\beta_j} \\
&= -\frac{1}{(p-1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j} + \frac{p}{(p-1)}\frac{\sum_{j=1}^{m}\beta_j b_{i(j)}}{\sum_{j=1}^{m}\beta_j}.
\end{aligned}$$

Then,

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p-1)\sum_{j=1}^{m}\beta_j}. \quad \text{(D.9)}$$

On the other side, from Equation (D.3), we can get the estimate of the common fraction p as the following way:

$$\begin{aligned}
\frac{\partial \ln L}{\partial p} &= \frac{2v}{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) = 0 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}\alpha_j + \sum_{j=1}^{m}\beta_j\frac{(b_{i(j)} - pS_i - \mu_{Hi})}{\sigma_i}\frac{S_i}{\sigma_i} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i}{\sigma_i^2}b_{i(j)} - p\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i^2}{\sigma_i^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{Hi}\frac{\beta_j S_i}{\sigma_i^2}.
\end{aligned}$$

Thereby,

$$\hat{p} \quad = \quad \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i}{\sigma_i^2}b_{i(j)} - \sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{Hi}\frac{\beta_j S_i}{\sigma_i^2}}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i^2}{\sigma_i^2}}.$$

We can also write the above equation as below:

$$\begin{aligned}
&\hat{p}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i^2}{\sigma_i^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i}{\sigma_i^2}b_{i(j)} + \sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{Hi}\frac{\beta_j S_i}{\sigma_i^2} = 0 \\
&\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i}{\sigma_i^2}b_{i(j)} - \hat{p}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j S_i^2}{\sigma_i^2} - \sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{Hi}\frac{\beta_j S_i}{\sigma_i^2} = 0. \qquad \text{(D.10)}
\end{aligned}$$

Then, by substituting Equations (D.8) and (D.9) into Equation (D.10):

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}b_{i(j)}\left[\frac{-\sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)}}{(p-1)\sum_{j=1}^{m}\beta_j}\right] - p\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}\left[\frac{-\sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)}}{(p-1)\sum_{j=1}^{m}\beta_j}\right]^2 \\
-\ &\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\beta_j}{\sigma_i^2}\left[\frac{p\sum_{j=1}^{m}\beta_j(a_{ij} - \sum_{j=1}^{m}\beta_j(b_{ij})}{(p^2+1)\sum_{j=1}^{m}\beta_j}\right] \\
\times\ &\left[\frac{-\sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)}}{(p-1)\sum_{j=1}^{m}\beta_j}\right] = 0,
\end{aligned}$$

and by solving this expression, we find zero equalities in both sides.

Finally, from Equation (D.5), we can obtain MLE of σ for each gene i by setting this expres-

sion to zero via:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \sigma_i} &= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m} g(z_{b_{i(j)}}) = 0 \\
&= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\frac{2m}{\sigma_i} &= \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\sigma_i &= \frac{v}{km}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\frac{km\sigma_i}{v} &= \sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \sum_{j=1}^{m}\left(\alpha_j + \beta_j\frac{a_{i(j)} - S_i - \mu_{Hi}}{\sigma_i} + \alpha_j + \beta_j\frac{b_{i(j)} - pS_i - \mu_{Hi}}{\sigma_i}\right) \\
\frac{km\sigma_i^2}{v} &= \sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j S_i - \sum_{j=1}^{m}\beta_j\mu_{Hi} + \sum_{j=1}^{m}\beta_j b_{i(j)} \\
&- \sum_{j=1}^{m}\beta_j pS_i - \sum_{j=1}^{m}\beta_j\mu_{Hi} \\
&= \sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)}) - (p+1)S_i\sum_{j=1}^{m}\beta_j - 2\mu_{Hi}\sum_{j=1}^{m}\beta_j.
\end{aligned}$$

Hence,

$$\hat{\sigma_i}^2 = \frac{\sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)}) - (p+1)S_i\sum_{j=1}^{m}\beta_j - 2\mu_{Hi}\sum_{j=1}^{m}\beta_j}{\frac{km}{v}}.$$

By substituting Equations (D.8) and (D.9) into the above equation, we can derive:

$$\begin{aligned}
\frac{km\hat{\sigma}_i^2}{v} &= \sum_{j=1}^{m}\beta_j(a_{i(j)}+b_{i(j)}) - (p+1)\left[\frac{-\sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)}}{(p-1)\sum_{j=1}^{m}\beta_j}\right] \\
&- 2\sum_{j=1}^{m}\beta_j\left[\frac{p\sum_{j=1}^{m}\beta_j(a_{ij} - \sum_{j=1}^{m}\beta_j(b_{ij})}{(p^2+1)\sum_{j=1}^{m}\beta_j}\right] \\
&= \sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)} \\
&- \frac{p+1}{p-1}\sum_{j=1}^{m}\beta_j b_{i(j)} + \frac{p+1}{p-1}\sum_{j=1}^{m}\beta_j a_{i(j)} - \frac{2p}{p-1}\sum_{j=1}^{m}\beta_j a_{i(j)} \\
&+ \frac{p+1}{p-1}\sum_{j=1}^{m}\beta_j a_{i(j)} - \frac{p}{p-1}\sum_{j=1}^{m}\beta_j a_{i(j)} \\
&= \left[1 + \frac{p+1}{p-1} - \frac{2p}{p-1}\right]\sum_{j=1}^{m}\beta_j a_{i(j)} + \left[1 - \frac{p+1}{p-1} + \frac{2}{p-1}\right]\sum_{j=1}^{m}\beta_j b_{i(j)},
\end{aligned}$$

since

$$1 + \frac{p+1}{p-1} - \frac{2p}{p-1} = 0 \text{ and } 1 - \frac{p+1}{p-1} + \frac{2}{p-1} = 0.$$

Then, we get the estimate of $\hat{\sigma}_i$ as below:

$$\frac{km\hat{\sigma}_i^2}{v} = 0.$$

Hereby,

$$\hat{\sigma}_i = 0,$$

which implies an infeasible estimate for σ_i.

APPENDIX E

DERIVATION of ESTIMATORS of ALTERNATIVE MODEL 4

In modelling perfect matches PM and mismatches MM intensities for each probe j and gene i, we assume the following relation under long-tailed symmetric (LTS) distribution on the log-scale:

$$\begin{aligned} \mathrm{PM}_{ij} &= a_{ij} \sim \mathrm{LTS}(S_i + \mu_{Hi}, \sigma_i^2) \\ \mathrm{MM}_{ij} &= b_{ij} \sim \mathrm{LTS}(p_i S_i + \mu_{Hi}, \sigma_i^2), \end{aligned} \tag{E.1}$$

for gene specific true signal S_i, background intensity μ_{Hi}, and variance σ_i^2.

Thereby, the associated likelihood can be written as follows:

$$\begin{aligned} L(S_i, \mu_{Hi}, p_i \mid a, b) &= \prod_{i=1}^{n}\prod_{j=1}^{m} f(a_{ij}) f(b_{ij}) \\ L &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(a_{ij}-S_i-\mu_{Hi})^2}{2\sigma_i^2}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(b_{ij}-p_i S_i-\mu_{Hi})^2}{2\sigma_i^2}} \end{aligned}$$

under $a = (a_{11}, \ldots, a_{ij}, \ldots, a_{nm})$ and $b = (b_{11}, \ldots, b_{ij}, \ldots, b_{nm})$, which is proportional to

$$L \propto \prod_{i=1}^{n}\prod_{j=1}^{m} (\frac{1}{\sigma_i})(1 + \frac{z_{a_{ij}}^2}{k})^{-v} \prod_{i=1}^{n}\prod_{j=1}^{m} (\frac{1}{\sigma_i})(1 + \frac{z_{b_{ij}}^2}{k})^{-v},$$

for the shape parameter $v \geq 2$ and $k = 2v - 3$. Moreover,

$$z_{a_{ij}} = \frac{(a_{ij} - S_i - \mu_{Hi})}{\sigma_i} \text{ and } z_{b_{ij}} = \frac{(b_{ij} - p_i S_i - \mu_{Hi})}{\sigma_i}.$$

Then, the logarithm of L can be derived as:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right]+\sum_{i=1}^{n}\sum_{j=1}^{m}\ln\left[\left(\frac{1}{\sigma_i}\right)\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i)+\ln\left(1+\frac{z_{a_{ij}}^2}{k}\right)^{-v}\right]+\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(-\ln\sigma_i)+\ln\left(1+\frac{z_{b_{ij}}^2}{k}\right)^{-v}\right] \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i)+v\ln\left(1+\frac{z_{a_{ij}}^2}{k}\right)\right]-\sum_{i=1}^{n}\sum_{j=1}^{m}\left[(\ln\sigma_i)+v\ln\left(1+\frac{z_{b_{ij}}^2}{k}\right)\right].
\end{aligned}$$

In order to get the MLE of model parameters, we take the first derivatives of $\ln L$ with respect to each parameter as below:

$$\begin{aligned}
\frac{\partial\ln L}{\partial\mu_{Hi}} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{j=1}^{m}\frac{(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial p_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{S_i(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial S_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^2\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}}+\frac{2v}{k}\sum_{j=1}^{m}\frac{-p_i(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^2\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}} \\
\frac{\partial\ln L}{\partial\sigma_i} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right] \\
&+ \sum_{i=1}^{n}\sum_{j=1}^{m}\left[-\frac{1}{\sigma_i}+\frac{(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i^3(1+\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2})}\frac{2v}{k}\right],
\end{aligned}$$

where

$$\begin{aligned}
z_{a_{ij}} &= \frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i} \\
z_{b_{ij}} &= \frac{(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i}.
\end{aligned}$$

Then, by approximating the common nonlinear functions $g(z)=\frac{z}{1+\frac{z^2}{k}}$ for PM and MM via:

$$g(z_{a_{ij}})=\frac{\frac{(a_{ij}-S_i-\mu_{Hi})}{\sigma_i}}{1+\frac{(a_{ij}-S_i-\mu_{Hi})^2}{k\sigma_i^2}} \text{ and } g(z_{b_{ij}})=\frac{\frac{(b_{ij}-p_iS_i-\mu_{Hi})}{\sigma_i}}{1+\frac{(b_{ij}-p_iS_i-\mu_{Hi})^2}{k\sigma_i^2}},$$

respectively, under the first order Taylor expansions as:

$$g(z_{a_{i(j)}})=\alpha_j+\beta_j z_{a_{i(j)}} \text{ and } g(z_{b_{i(j)}})=\alpha_j+\beta_j z_{b_{i(j)}},$$

as well as

$$\alpha_j = \frac{\frac{2t_j^3}{k}}{(1+\frac{t_j^2}{k})} \text{ and } \beta_j = \frac{(1-\frac{t_j^2}{k})}{(1+\frac{t_j^2}{k})^2},$$

we can present the partial derivatives of MLE as follows:

$$\frac{\partial \ln L}{\partial \mu_{Hi}} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \tag{E.2}$$

$$\frac{\partial \ln L}{\partial p_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \tag{E.3}$$

$$\frac{\partial \ln L}{\partial S_i} = \frac{2v}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \frac{2vp_i}{k}\sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) \tag{E.4}$$

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}g(z_{b_{i(j)}}), \tag{E.5}$$

for the probe based ordered and standardized PM and MM (denoted by $z_{a_{i(j)}}$ and $z_{b_{i(j)}}$, respectively) values and the jth quantile of the student-t distribution(denoted by $t_(j)$).

Finally, to obtain the estimates of parameters, we set Equations (E.2) - (E.5) to zero. Thereby, from Equation (E.2):

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_{Hi}} &= \sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{a_{i(j)}}) + \sum_{j=1}^{m}\frac{1}{\sigma_i}g(z_{b_{i(j)}}) = 0 \\
&= \sum_{j=1}^{m}\frac{1}{\sigma_i}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \frac{1}{\sigma_i}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \sum_{j=1}^{m}(\alpha_j + \beta_j z_{a_{i(j)}} + \alpha_j + \beta_j z_{b_{i(j)}}) \\
&= \sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\beta_j z_{a_{i(j)}}\sigma_i + \sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\beta_j z_{b_{i(j)}} \\
&= \sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\beta_j z_{a_{i(j)}} + \sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\beta_j z_{b_{i(j)}},
\end{aligned}$$

by taking $\sum_{j=1}^{m} \alpha_j = 0$ due to the symmetry. Then,

$$\begin{aligned}
\frac{\partial \ln L}{\partial \mu_{Hi}} &= \sum_{j=1}^{m} \beta_j z_{a_{i(j)}} + \sum_{j=1}^{m} \beta_j z_{b_{i(j)}} = 0 \\
&= \sum_{j=1}^{m} \beta_j \frac{(a_{i(j)} - S_i - \mu_{Hi})}{\sigma_i} + \sum_{j=1}^{m} \beta_j \frac{(a_{i(j)} - p_i S_i - \mu_{Hi})}{\sigma_i} \\
&= \sum_{j=1}^{m} \beta_j (a_{i(j)} - S_i - \mu_{Hi}) + \sum_{j=1}^{m} \beta_j (a_{i(j)} - p_i S_i - \mu_{Hi}).
\end{aligned}$$

From the above equations we get the following expression:

$$\begin{aligned}
\sum_{j=1}^{m} \beta_j a_{i(j)} - \sum_{j=1}^{m} \beta_j S_i - \mu_{Hi} \sum_{j=1}^{m} \beta_j &= -\sum_{j=1}^{m} \beta_j b_{i(j)} + p_i \sum_{j=1}^{m} \beta_j S_i + \mu_{Hi} \sum_{j=1}^{m} \beta_j \\
2\mu_{Hi} \sum_{j=1}^{m} \beta_j &= \sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p_i + 1) S_i \sum_{j=1}^{m} \beta_j.
\end{aligned}$$

As a result, $\hat{\mu}_H$ is found as:

$$\hat{\mu}_{Hi} = \frac{\sum_{j=1}^{m} \beta_j (a_{i(j)} + b_{i(j)}) - (p_i + 1) S_i \sum_{j=1}^{m} \beta_j}{2 \sum_{j=1}^{m} \beta_j}. \tag{E.6}$$

On the other hand, from Equation (E.4):

$$\begin{aligned}
\frac{\partial \ln L}{\partial S_i} &= \frac{2v}{k} \sum_{j=1}^{m} \frac{1}{\sigma_i} g(z_{a_{i(j)}}) + \frac{2v p_i}{k} \sum_{j=1}^{m} \frac{1}{\sigma_i} g(z_{b_{i(j)}}) = 0 \\
&= \frac{2v}{k} \frac{1}{\sigma_i} \left[\sum_{j=1}^{m} (g(z_{a_{i(j)}}) + p_i g(z_{b_{i(j)}})) \right] \\
&= \frac{2v}{k\sigma_i} \sum_{j=1}^{m} ((\alpha_j + \beta_j z_{a_{i(j)}}) + \frac{2v}{k\sigma_i} \sum_{j=1}^{m} ((\alpha_j + \beta_j z_{b_{i(j)}}) \\
&= \frac{1}{\sigma_i^2} \sum_{j=1}^{m} \beta_j (a_{i(j)} - S_i - \mu_{Hi}) + \frac{p_i}{\sigma_i^2} \sum_{j=1}^{m} \beta_j (b_{i(j)} - p_i S_i - \mu_{Hi}) \\
&= \sum_{j=1}^{m} \beta_j a_{i(j)} - \sum_{j=1}^{m} \beta_j S_i - \mu_{Hi} \sum_{j=1}^{m} \beta_j + p_i \sum_{j=1}^{m} \beta_j b_{i(j)} \\
&- p_i^2 \sum_{j=1}^{m} \beta_j S_i - p_i \mu_{Hi} \sum_{j=1}^{m} \beta_j.
\end{aligned}$$

So,

$$(p_i^2 + 1) S_i \sum_{j=1}^{m} \beta_j = \sum_{j=1}^{m} \beta_j a_{i(j)} + p_i \sum_{j=1}^{m} \beta_j b_{i(j)} - (p_i + 1) \mu_{Hi} \sum_{j=1}^{m} \beta_j,$$

the estimate of the true signal can be found by:

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - (p_i+1)\mu_{Hi}\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}. \tag{E.7}$$

But by substituting Equation E.7 into Equation E.6, we can also write $\hat{\mu}_H$ as the following form:

$$\begin{aligned}\hat{\mu}_{Hi} &= \frac{\sum_{j=1}^{m}\beta_j(a_{i(j)}+b_{i(j)})}{2\sum_{j=1}^{m}\beta_j} \\ &- \frac{(p_i+1)\left[\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}+p_i\sum_{j=1}^{m}\beta_j b_{i(j)}-(p_i+1)\mu_{Hi}\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j}{2\sum_{j=1}^{m}\beta_j}.\end{aligned}$$

Accordingly,

$$\hat{\mu}_{Hi} = \frac{p_i\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{j=1}^{m}\beta_j b_{(ij)}}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}. \tag{E.8}$$

Similarly, by substituting Equation E.8 into Equation E.7, an alternative form of $\hat{S}_i$ can be found as:

$$\begin{aligned}\hat{S}_i &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)} - (p_i+1)\mu_{Hi}\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j} \\ &= \frac{\sum_{j=1}^{m}\beta_j a_{i(j)} + p_i\sum_{j=1}^{m}\beta_j b_{i(j)}}{(p_i^2+1)\sum_{j=1}^{m}\beta_j} \\ &- \frac{(p_i+1)\left[\frac{p_i\sum_{j=1}^{m}\beta_j a_{(ij)}-\sum_{j=1}^{m}\beta_j b_{(ij)}}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j}{(p_i^2+1)\sum_{j=1}^{m}\beta_j} \\ &= -\frac{1}{(p_i-1)}\frac{\sum_{j=1}^{m}\beta_j a_{i(j)}}{\sum_{j=1}^{m}\beta_j} + \frac{p_i}{(p_i-1)}\frac{\sum_{j=1}^{m}\beta_j b_{i(j)}}{\sum_{j=1}^{m}\beta_j}.\end{aligned}$$

Thus,

$$\hat{S}_i = \frac{\sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j a_{i(j)}}{(p_i-1)\sum_{j=1}^{m}\beta_j}. \tag{E.9}$$

On the other side, from Equation (E.3), we can get the estimate of the common fraction p_i as

below:

$$\begin{aligned}
\frac{\partial \ln L}{\partial p_i} &= \frac{2v}{k}\sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) = 0 \\
&= \sum_{j=1}^{m}\frac{S_i}{\sigma_i}g(z_{b_{i(j)}}) \\
&= \frac{S_i}{\sigma_i}\sum_{j=1}^{m}g(z_{b_{i(j)}}) \\
&= \sum_{j=1}^{m}g(z_{b_{i(j)}}) \\
&= \sum_{j=1}^{m}\alpha_j + \sum_{j=1}^{m}\beta_j\frac{(b_{i(j)} - p_iS_i - \mu_{Hi})}{\sigma_i} \\
&= \sum_{j=1}^{m}\beta_j\frac{(b_{i(j)} - p_iS_i - \mu_{Hi})}{\sigma_i} \\
&= \sum_{j=1}^{m}\beta_j(b_{i(j)} - p_iS_i - \mu_{Hi}) \\
&= \sum_{j=1}^{m}\beta_j b_{i(j)} - \sum_{j=1}^{m}\beta_j p_i S_i - \sum_{j=1}^{m}\beta_j\mu_{Hi}).
\end{aligned}$$

Hence,

$$\hat{p} = \frac{\sum_{j=1}^{m} b_{i(j)} - \mu_{Hi}\sum_{j=1}^{m}\beta_j}{S_i\sum_{j=1}^{m}\beta_j}.$$

We can write the above equation as follows, too:

$$\hat{p}S_i\sum_{j=1}^{m}\beta_j - \sum_{j=1}^{m}b_{i(j)} + \mu_{Hi}\sum_{j=1}^{m}\beta_j = 0. \tag{E.10}$$

Then, by substituting Equations (E.8) and (E.9) into Equation (E.10):

$$\begin{aligned}
0 &= \hat{p}_i\left[\frac{-\sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)}}{(p_i - 1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j - \sum_{j=1}^{m}b_{i(j)} \\
&+ \left[\frac{p_i\sum_{j=1}^{m}\beta_j a_{(ij)} - \sum_{j=1}^{m}\beta_j b_{(ij)}}{(p_i^2 + 1)\sum_{j=1}^{m}\beta_j}\right]\sum_{j=1}^{m}\beta_j
\end{aligned}$$

and by solving the above equation, we get:

$$\left[\frac{p_i}{p_i - 1} - 1 - \frac{1}{p_i - 1}\right]\sum_{j=1}^{m}\beta_j b_{(ij)} = 0,$$

while

$$\left[\frac{p_i}{p_i - 1} - 1 - \frac{1}{p_i - 1}\right] = 0.$$

Above equation gives us no solution, which implies an infeasible estimate for p_i.

Finally, from Equation (E.5), we can obtain MLE of σ for each gene i by setting this expression to zero via:

$$\begin{aligned}
\frac{\partial \ln L}{\partial \sigma_i} &= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m} g(z_{a_{i(j)}}) + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m} g(z_{b_{i(j)}}) \\
&= -\frac{2m}{\sigma_i} + \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) = 0 \\
\frac{2m}{\sigma_i} &= \frac{2v}{k\sigma_i^2}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\sigma_i &= \frac{v}{km}\sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
\frac{km\sigma_i}{v} &= \sum_{j=1}^{m}(g(z_{a_{i(j)}}) + g(z_{b_{i(j)}})) \\
&= \sum_{j=1}^{m}\left(\alpha_j + \beta_j\frac{a_{i(j)} - S_i - \mu_{Hi}}{\sigma_i} + \alpha_j + \beta_j\frac{b_{i(j)} - p_iS_i - \mu_{Hi}}{\sigma_i}\right) \\
\frac{km\sigma_i^2}{v} &= \sum_{j=1}^{m}\beta_j a_{i(j)} - \sum_{j=1}^{m}\beta_j S_i - \sum_{j=1}^{m}\beta_j\mu_{Hi} + \sum_{j=1}^{m}\beta_j b_{i(j)} \\
&- \sum_{j=1}^{m}\beta_j p_i S_i - \sum_{j=1}^{m}\beta_j\mu_{Hi} \\
&= \sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)}) - (p_i + 1)S_i\sum_{j=1}^{m}\beta_j - 2\mu_{Hi}\sum_{j=1}^{m}\beta_j.
\end{aligned}$$

Thereby,

$$\hat{\sigma_i}^2 = \frac{km}{v}\left(\sum_{j=1}^{m}\beta_j(a_{i(j)} + b_{i(j)}) - (p_i + 1)S_i\sum_{j=1}^{m}\beta_j - 2\mu_{Hi}\sum_{j=1}^{m}\beta_j\right).$$

In the end, by substituting Equations (E.8) and (E.9) into the above equation, we get the most simple form of σ_i, which are the non-linear functions of p_i, resulting in no explicit expressions

for the model parameter of Equation (E.1):

$$\begin{aligned}
\frac{km\hat{\sigma}_i^2}{v} &= \sum_{j=1}^{m}\beta_j(a_{i(j)}+b_{i(j)}) - (p_i+1)\left[\frac{-\sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)}}{(p_i-1)\sum_{j=1}^{m}\beta_j}\right] \\
&- 2\sum_{j=1}^{m}\beta_j\left[\frac{p_i\sum_{j=1}^{m}\beta_j(a_{(ij)} - \sum_{j=1}^{m}\beta_j(b_{(ij)})}{(p_i^2+1)\sum_{j=1}^{m}\beta_j}\right] \\
&= \sum_{j=1}^{m}\beta_j a_{i(j)} + \sum_{j=1}^{m}\beta_j b_{i(j)} - \frac{p_i+1}{p_i-1}\sum_{j=1}^{m}\beta_j b_{i(j)} + \frac{p_i+1}{p_i-1}\sum_{j=1}^{m}\beta_j a_{i(j)} \\
&- \frac{2p_i}{p_i-1}\sum_{j=1}^{m}\beta_j a_{i(j)} + \frac{p_i+1}{p_i-1}\sum_{j=1}^{m}\beta_j a_{i(j)} - \frac{p_i}{p_i-1}\sum_{j=1}^{m}\beta_j a_{i(j)} \\
&= \left[1 + \frac{p_i+1}{p_i-1} - \frac{2p_i}{p_i-1}\right]\sum_{j=1}^{m}\beta_j a_{i(j)} + \left[1 - \frac{p_i+1}{p_i-1} + \frac{2}{p_i-1}\right]\sum_{j=1}^{m}\beta_j b_{i(j)},
\end{aligned}$$

since

$$1 + \frac{p_i+1}{p_i-1} - \frac{2p_i}{p_i-1} = 0 \text{ and } 1 - \frac{p_i+1}{p_i-1} + \frac{2}{p_i-1} = 0.$$

Then we obtain 0 as the estimate of the variance term σ_i as below:

$$\begin{aligned}
\frac{km\hat{\sigma}_i^2}{v} &= 0 \\
\hat{\sigma}_i &= 0,
\end{aligned}$$

which implies infeasible estimator for the standard deviation.

APPENDIX F

R CODES of the multi-RGX FUNCTION

Inputs values for the multi-RGX function :

PMs: Perfect matches values on the log-scale

MMs: Mismatches values on the log-scale

n.genes: Number of genes used in the analysis

n.probes: Number of probes for each gene

shape.par: Shape parameter of the long-tailed symmetric distribution

maximum-shape.par: Maximum number of shape parameters, which can be tested during the grid search of the optimal shape parameters. The default is NULL.

```
multi.rgx
<-function(PMs,MMs,n.genes,n.probes,shape.par,maximum_shape.par=NULL){

    if(is.null(maximum-shape.par)){
        max-shape.par <- 40
    }else{
        max-shape.par<- maximum-shape.par
    }
    L.all <- NULL
    muH.all<-NULL
    sigma.all<-NULL
    p.all<-NULL
    Si.all<-NULL
    PM.mat<-NULL
    MM.mat<-NULL
    alln.genes<-n.genes
    alln.probes<-n.probes
    Si<-rep(0,n.genes)
    shape.par.vector <- seq(2, max_shape.par, by = 0.5)
    for(i46 in 1:alln.genes){
        new.gene<-c((alln.probes*(i46-1)+1):(alln.probes*i46))
```

```
        PM.mat<-rbind(PM.mat,PMs[new.gene,])
        MM.mat<-rbind(MM.mat,MMs[new.gene,])
    }
    for (i6 in 1:length(shape.par.vector)){
        v <- shape.par.vector[i6]
        shape.par <- v
        k <- 2*v-3
        dof <- 2*v-1
        comp.weight<-alpha.beta(alln.probes,k,dof)
        myalphas<-comp.weight$comp.alphas
        mybetas<-comp.weight$comp.betas
        tmybetas<-t(mybetas)
        PMorder<-matrix(0,nrow=alln.genes,ncol=alln.probes)
        MMorder<-matrix(0,nrow=alln.genes,ncol=alln.probes)
        sum.21<-rep(0,length=alln.genes)
        sum.22<-rep(0,length=alln.genes)
        in.SSA<-rep(0,length=alln.genes)
        in.SSB<-rep(0,length=alln.genes)
        in.SSAB<-rep(0,length=alln.genes)
        B<-rep(0,length=alln.genes)
        C<-rep(0,length=alln.genes)
        sum21<-rep(0,length=alln.genes)
        sum22<-rep(0,length=alln.genes)

-----------------------------------------------------------------
Ordering perfect matches and mismatches for each gene, according to
their probes.
-----------------------------------------------------------------

        for (i3 in 1:alln.genes){
            PMorder[i3,] <- sort(PM.mat[i3,])
            MMorder[i3,] <- sort(MM.mat[i3,])
        }
        for(i4 in 1:alln.genes){
            sum21[i4] <- sum(mybetas*PMorder[i4,])
            sum22[i4] <- sum(mybetas*MMorder[i4,])
            in.SSA[i4] <- sum(mybetas)*(sum21[i4]
                        -(colMeans(PMorder)%*%mybetas))^2
            in.SSB[i4] <- sum(mybetas)*(sum22[i4]
                        -(colMeans(MMorder)%*%mybetas))^2
            in.SSAB[i4] <- sum(mybetas)*
                         (sum21[i4]-sum(colMeans(PMorder)%*%mybetas))%*%
                         (sum22[i4]-sum(colMeans(MMorder)%*%mybetas))
        }
        SSA<-sum(in.SSA)
        SSB<-sum(in.SSB)
        SSAB<-sum(in.SSAB)

---------------------------------------
```

```
Estimation of fraction, background signal, and true signal values,
respectively.
-----------------------------------------

        p.est <- ((SSB-SSA)+sqrt((SSA-SSB)^2+4*SSAB^2))/(2*SSAB)
        mu.H <- (p.est*sum(sum21)-sum(sum22))
              /(alln.genes*(p.est-1)*sum(mybetas))
        Si <- (sum21+p.est*sum22-(1+p.est)*mu.H*sum(mybetas))/
            ((1+p.est^2)*sum(mybetas))

-----------------------------------------------------
        for(i10 in 1:alln.genes){
            B[i10] <- sum(myalphas*(PMorder[i10,]-MMorder[i10,]))
            C[i10] <- sum(mybetas*(PMorder[i10,]-Si[i10]-mu.H)^2)+
                      sum(mybetas*(MMorder[i10,]-p.est*Si[i10]-mu.H)^2)
        }
        all.B<-sum(B)*v/k
        all.C<-sum(C)*v/k

--------------------------------
Estimation of standard deviation
--------------------------------

        sigma <- (all.B+sqrt(all.B^2+4*alln.genes*alln.probes*all.C))/
               (2*alln.genes*alln.probes-2)

---------------------------------------------------
Finding optimal shape parameters among alternatives
---------------------------------------------------

        in.L-A <- rep(0, length=alln.genes)
        in.L-B <- rep(0, length=alln.genes)
        L-constant<-(sqrt(k)*gamma(1/2)*gamma(v-1/2))/gamma(v)
        L-constant<-1/L-constant
        for (i5 in 1:alln.genes){
            in.L_A[i5] <- sum(log((1 + ((PM.mat[i5]-Si-mu.H)^2)/
                        (k*sigma^2)))^(-v))
            in.L_B[i5] <- sum(log((1 + ((MM.mat[i5]-p.est*Si-mu.H)^2)/
                        (k*sigma^2)))^(-v))
        }
        L-A <- sum(in.L-A)
        L-B <- sum(in.L-B)
        L.all <- c(L.all, (log(L_constant)-2*length(alln.genes)*
                 length(alln.probes)*log(sigma) + L-A + L-B))
    }
    max.L<-max(L.all)
    for(i11 in 1:length(shape.par.vector)){
        if(max.L==L.all[i11]){our.cell<-i11}
    }
```

```
-----------------------------------------------------------
Taking the final estimates for the optimal shape parameter
-----------------------------------------------------------

    p.est<-p.all[our.cell]
    sigma<-sigma.all[our.cell]
    mu.H<-muH.all[our.cell]
    Si<-Si.all[our.cell,]
    L<-L.all[our.cell]
    our.shape-par<-shape.par.vector[our.cell]
    return
}

-------------------------------------------------
Calculation of weight functions in MML estimators
-------------------------------------------------

alpha.beta <- function(ourn.probes=alln.probes,ourk=k,ourdof=dof){

    order.probe <- c(1:ourn.probes)
    quan.probe <- order.probe/(ourn.probes+1)
    tvalue.probe <- rep(0,length=ourn.probes)
    for(i34 in 1:ourn.probes){
        tvalue.probe[i34] <- qt(quan.probe[i34],ourdof)
                             *sqrt(ourk/ourdof)
    }
    beta1.probe <- (1-tvalue.probe[1]^2/ourk)/
                (1+tvalue.probe[1]^2/ourk)^2
    if(beta1.probe<0){
        alphas.probe <- (1/ourk*tvalue.probe^3)/
                        (1+tvalue.probe^2/ourk)^2
        betas.probe <- 1/(1+tvalue.probe^2/ourk)^2
    }else{
        alphas.probe <- (2*tvalue.probe^3/ourk)/
                        (1+tvalue.probe^2/ourk)^2
        betas.probe <- (1-tvalue.probe^2/ourk)/
                        (1+tvalue.probe^2/ourk)^2
    }
    output <-list(comp.alphas=alphas.probe,comp.betas=betas.probe,
                tvalue.probe=tvalue.probe)
    output
}
```