Research Article

Tülay Akal, Vilda Purutçuoğlu* and Gerhard-Wilhelm Weber

Robust background normalization method for one-channel microarrays
Tek-kanallı mikrodizinlerde sağlam ardalan normalizasyon metodu

DOI 10.1515/tjb-2016-0231
Received February 4, 2016; accepted June 23, 2016; previously published online January 20, 2017

Abstract

Background: Microarray technology, aims to measure the amount of changes in transcribed messages 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 signals. For this reason, we present a novel gene expression index, called multi-RGX (Multiple-probe Robust Gene Expression Index) for one-channel microarrays.

Methods: Multi-RGX, different from other gene expression indices, considers the long-tailed symmetric (LTS) density, covering a wider range of distributions for modelling gene expressions on the log-scale, resulting in robust inference and it takes into account both probe and gene specific intensities. Furthermore, we derive the covariance-variance matrix of model parameters from the observed Fisher information matrix and test the performance of the multi-RGX method in three different datasets.

Results: Our method is found to be a promising method regarding its alternatives in terms of accuracy and computational time.

Conclusion: Multi-RGX gives accurate results with respect to its alternatives, with a reduction in computational cost.

Keywords: Gene expression index; Oligonucleotide; Modified maximum likelihood approach; Observed Fisher information matrix; Bioinformatics.

Introduction

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 to measure the gene expressions. By collecting the amount of the mRNA transcripts measured, an approximation to the gene expression level, are turned into pictures [1].

The oligonucleotide is a type of single channel microarrays with a 25-base pair long and the Affymetrix Gene-Chip is an example of the popular oligonucleotides. In this array, we observe probes that 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 cRNA and the latter is aimed to measure the faulty signals on the arrays by changing the 13th base pair of PM.

In a microarray study, we can define three main sources of variations in signals, namely, the nonspecific hybridization, the background signal and the stray signal. They are also shortly called as the background errors. Accordingly, we call these three types of variations as the background errors [2, 3].

Thereby, the term gene expression index stands for the mathematical method to estimate the true expression level in the oligonucleotides under such erroneous signals. There are a number of indices developed to infer the signals. In this study, we propose a novel gene expression index, called multiple-probe robust gene expression index, multi-RGX, for modelling the observed gene expressions in oligonucleotides.

Among many choices, one of the most common gene expression indices is the MBEI method [4], which cannot make inference for large numbers of arrays as the calculation is based on the least square approach. On the other hand, MASS0.0 [5], suggesting a smart solution to solve the less estimated intensities in PM regarding the estimated MM, may cause bias in inference of the true signal. Similarly, RMA [6] also fails under this condition. Similar to RMA, mgMOS [7] also accepts that the MM probe intensities are treated as only coming from the background signal. But different from RMA, this model makes use of latent variables, standing for different binding affinities of probes within a specified probe set in order to describe the correlation between the PM and MM intensities.

On the other side, GC-RMA [8] is the extended version of RMA by considering the GC content in the data and by adding, the first time, the presence of true signals in the MM probes. On the contrary, in practice, it equates the fraction of the true signal in MM to zero in order to simplify the calculation. But as the computation in the Bayesian framework is demanding, multi-mgMOS [9] considers performing it via a maximum likelihood approximation under the same model of BGX [10] so that the computational cost can reduce. However, being more efficient than BGX, the calculation is still time demanding for large datasets due to the application of the Bayesian method. Accordingly, FGX [2] is proposed to decrease the underlying cost by fully using the maximum likelihood approach in inference of model parameters. But, it is based on a strict normality assumption and the estimates are found for the probe mean of each gene.

Another recent method is called the Hook calibration method [11] which is a singlechip calibration alternative taking mismatched probes as the internal reference. Also, RMX [12] tries to find optimal robust estimators for multiple raw signals by making use of the Kolmogorov distance which is non-parametric. On the other hand, Lahti et al. [13] propose an online learning algorithm, called RPA which gives advantage to update probe-level parameters when the observations comes up. However, in modeling, it still uses the RMA method resulting in the same drawback in the description of the MM probes.

Hereby, in this study, we suggest a new gene expression index, which includes both gene and probe specific intensities, alike FGX and RGX, to model true signals in oligonucleotides. In this method, we again use the long-tailed symmetric family (LTS) like RGX in order to keep the robustness. The robust estimators are less sensitive to outliers in the data and tolerate departures from target distributions compared to RGX. In inference of the model parameters, we implement the modified maximum likelihood method, and get explicit and close form for each estimator. Moreover, we derive the variances and covariances of model parameters via the observed Fisher information matrix. From the analyses by different datasets, we conclude that the multi-RGX method is promising in terms of accuracy and computational demand and can be a helpful tool for the application in computational biology, bioinformatics and biomolecular engineering.

Materials and methods

Multiple-probe robust gene expression index: multi-RGX

As described in the previous section, the major aim of our study is to suggest a new gene expression index. Moreover, we aim that the new index can estimate the true signals more accurately, since we model not only gene specific intensities, but also both gene and probe intensities. Finally, multi-RGX covers a wider range of distributions for the signals, hence, does not ignore any information in the data. Accordingly, it suggests a robust method to model the true signals.
Hereby, in this novel model, called as multi-RGX (multiple-probe robust gene expression index), we assume that the intensities of the PM and MM probes on the logarithmic scale have the density from the LTS, similar to the idea of RGX. On the other hand, different from this, it does not follow an iterative process in the optimization of the likelihood function, rather, it works with the explicit estimates of the desired parameters.

In a standard Affymetrix chip, accepting that there are \( m \) probes, and in each probe, there exist \( n \) genes coming from the PM and MM probes, it is assumed that both PM and MM have true signals in such a way that PM covers the complete true signals, whereas, MM includes only a fraction of them. To define the amount of the underlying true signals in MM, a fraction term \( p \) is defined. Hereby, the intensities coming from the PM probes are distributed as LTS with a mean containing a constant term and another term, which stands for the true signal changing per gene. Additionally, its variance is constant over genes and probes. Indeed, this assumption may be considered strong in modelling. However, from the application of this assumption in FGX and RGX indices, it has been found that it is competitive in terms of the accuracies of the signals with respect to the gene specific variance as BGX implements. Moreover, from alternative modellings via the gene specific variance, it has been observed that none of the estimators can have explicit solutions under a gene specific variance [14], and the estimates can be obtained via optimization steps which cause an increase in the computational time. On the other hand, different from PM, MM has a mean containing the same constant term and the part of the true signal. Hence, in modelling of the PM and MM intensities for each probe \( j (j=1, \ldots, m) \) and each gene \( i (i=1, \ldots, n) \), we consider the following distributional assumption on the log-scale:

\[
\text{PM}_i \sim \text{LTS}(S_i + \mu_i, \sigma^2) \quad \text{and} \quad \text{MM}_i \sim \text{LTS}(pS_i + \mu_i, \sigma^2),
\]

where \( S_i \) shows the true signal for the gene \( i \) and \( \mu_i \) denotes the constant background signal. Furthermore, \( \sigma^2 \) indicates the constant variance and \( p \) stands for the fraction value within the range of zero to one that implies the ratio of the true signal in the MM probes. Thereby, the corresponding likelihood of the model in Equation (1) is found via

\[
L(S_i, \mu_i, p|a, b) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(\text{PM}_i) f(\text{MM}_i),
\]

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{m} \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\text{PM}_i - S_i - \mu_i)^2}{2\sigma^2}} \right) \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\text{MM}_i - pS_i - \mu_i)^2}{2\sigma^2}} \right),
\]

which is proportional to

\[
L = \left( \frac{1}{\sigma} \right)^{n} \prod_{i=1}^{n} \prod_{j=1}^{m} \left( 1 + \frac{z_{\text{PM}}^2}{k} \right)^{\frac{-1}{2}} \left( 1 + \frac{z_{\text{MM}}^2}{k} \right)^{\frac{-1}{2}},
\]

where \( \nu \geq 4 \) and \( k = 2\nu - 3 \).

Then, the first derivatives of the function \( \ln L \) are taken with respect to each parameter as stated below:

\[
\frac{\partial \ln L}{\partial \mu_i} = \frac{2\nu}{k\sigma} \sum_{j=1}^{m} \frac{(\text{PM}_i - S_i - \mu_i)}{(\text{PM}_i - S_i - \mu_i)^2} 
+ \frac{2\nu}{k\sigma} \sum_{j=1}^{m} \frac{(\text{MM}_i - pS_i - \mu_i)}{(\text{MM}_i - pS_i - \mu_i)^2},
\]

\[
\frac{\partial \ln L}{\partial \sigma} = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[ \frac{-1}{\sigma^3} \left( 1 + \frac{z_{\text{PM}}^2}{k} \right)^{\frac{-1}{2}} \right] \left[ \frac{1}{\sigma^3} \left( 1 + \frac{z_{\text{MM}}^2}{k} \right)^{\frac{-1}{2}} \right],
\]

But it is observed that the underlying partial derivatives cannot produce unique solutions due to the repeated nonlinear functions in their formulas. Therefore, we suggest to use the modified maximum likelihood approach (MMLE) [15] that is based on approximating the nonlinear functions via the first order Taylor expansions around the \( i \)th population quantile \( t_{i(0)} \) of the student-t distribution with the \( (2\nu - 1) \) degrees of freedom. In the calculation, we apply the order statistics and define the partial derivatives of the objective log-likelihood function in Equation (2) in terms of small linear sub-pieces. Moreover, it is found that the estimators based on MMLE are asymptotically efficient and equivalent to MLE [15, 16].
In this study, we observe that the nonlinearity in the partial derivatives is caused by the following expressions such that

\[ z_{PMi} = \frac{(PM_i - S_i - \mu_H)}{\sigma} \quad \text{and} \quad z_{MMi} = \frac{(MM_i - pS_i - \mu_H)}{\sigma}. \]

Then, by approximating the common nonlinear functions

\[ g(z) = \frac{z}{1 + \frac{z}{k}} \quad \text{for PM as} \]

\[ g(z_{PMi}) = \frac{(PM_i - S_i - \mu_H)}{\sigma (1 + \frac{(PM_i - S_i - \mu_H)^2}{k\sigma^2})}, \]

and for MM as

\[ g(z_{MMi}) = \frac{(MM_i - pS_i - \mu_H)}{\sigma (1 + \frac{(MM_i - pS_i - \mu_H)^2}{k\sigma^2})}, \]

their first-order Taylor expansions can be derived by

\[ g(z_{PMi}) = \alpha_j + \beta_j z_{PMi}, \quad \text{and} \quad g(z_{MMi}) = \alpha_j + \beta_j z_{MMi}, \]

in which

\[ \alpha_j = \frac{2t_j}{k} \quad \text{and} \quad \beta_j = \left( \frac{1 - t_j}{k} \right)^{1/2}. \]

In these expressions, \( z_{PMi} \) and \( z_{MMi} \) show the ordered probes in increasing magnitude for each gene \( i \) in the PM and MM standardized intensities, respectively. Accordingly, \( g(z_{PMi}) \) and \( g(z_{MMi}) \) are their associated linearized functions via the Taylor series. Thereby, the partial derivatives of MMLE are as follows:

\[ \frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k\sigma} \sum_{j=1}^{m} \sum_{i=1}^{n} g(z_{PMi}) + \frac{2v}{k\sigma} \sum_{j=1}^{m} \sum_{i=1}^{n} g(z_{MMi}), \]

\[ \frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{m} g(z_{PMi}) + \frac{2v}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{m} g(z_{MMi}), \]

\[ \frac{\partial \ln L}{\partial \mu_H} = \frac{2v}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{m} g(z_{PMi}) + \frac{2v}{k\sigma} \sum_{i=1}^{n} \sum_{j=1}^{m} g(z_{MMi}), \]

Finally, by setting Equations (7)–(10) to zero, we get the optimal MML estimates of parameters.

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

\[ \hat{\mu}_H = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} (PM_{ij} - MM_{ij}) - (p+1) \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j S_{ij}}{2n \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j}. \]

(11)

On the other side, from Equation (9), we obtain the estimate of the true signal for each gene \( i \) as:

\[ \hat{S}_i = \frac{\sum_{j=1}^{m} \beta_j PM_{ij} + \sum_{j=1}^{m} \beta_j MM_{ij} - (p+1) \mu_H \sum_{j=1}^{m} \beta_j}{(p^2+1) \sum_{j=1}^{m} \beta_j}. \]

(12)

However, by substituting Equation (12) into Equation (11), we can define \( \hat{\mu}_H \) in an alternative way as follows:

\[ \hat{\mu}_H = \frac{p \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j PM_{ij} - \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j MM_{ij}}{np - \sum_{j=1}^{m} \beta_j}. \]

\[ \hat{\mu}_H = \frac{p \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j PM_{ij} - \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j MM_{ij}}{np - \sum_{j=1}^{m} \beta_j}. \]

(14)

where \( PM_{ij} = \sum_{j=1}^{m} \frac{PM_{ij}}{n} \) and \( MM_{ij} = \sum_{j=1}^{m} \frac{MM_{ij}}{n} \).

The estimate of the common fraction term \( p \) is computed by four roots and the optimal one which maximizes the log-likelihoods is found as the following equation:

\[ \hat{p} = \frac{(SS_{MM} - SS_{PM}) + \sqrt{(SS_{PM} - SS_{MM})^2 + 4SS_{PM,MM}^2}}{2SS_{PM,MM}}. \]

where

\[ SS_{PM} = \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \left[ \sum_{j=1}^{m} \beta_j PM_{ij} \right]^2, \]

\[ SS_{MM} = \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \left[ \sum_{j=1}^{m} \beta_j MM_{ij} \right]^2, \]

\[ SS_{PM,MM} = \sum_{j=1}^{m} \sum_{i=1}^{n} \beta_j \left[ \left( \sum_{j=1}^{m} \beta_j PM_{ij} \right) - \sum_{j=1}^{m} \beta_j \bar{PM} \right] \times \left[ \left( \sum_{j=1}^{m} \beta_j MM_{ij} \right) - \sum_{j=1}^{m} \beta_j \bar{MM} \right] \]

Finally, from Equation (10), we can get MMLE of \( \sigma \) by:

\[ \hat{\sigma} = B + \sqrt{B^2 + 4nmC}, \]

Adjusting the degree of freedom as \( 2nm - (n+2) \):
\[ \hat{\sigma} = \frac{B + \sqrt{B^2 + 4nmC}}{nm - n - 2}, \]

in which

\[ B = \frac{v}{k} \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j (P_{M(i)} - \hat{S}_i - \hat{\mu}_i)^2 \right], \]
\[ C = \frac{v}{k} \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j (P_{M(i)} - \hat{S}_i - \hat{\mu}_i)^2 \right]. \]

**Observed Fisher information matrix and estimators for variances and covariances**

The variance of the asymptotic maximum likelihood estimator can be found from the Fisher information matrix [17] and as stated previously that MMLE is asymptotically equivalent to it, resulting in the same feature of estimators.

Accordingly, the Fisher information matrix \( I \) is generated from the second partial derivatives of the loglikelihood function with respect to each parameter as stated below:

\[ I_1 = \frac{\partial^2 l}{\partial \mu_i \partial p} = - \frac{1}{\sigma^2}, \]
\[ I_2 = \frac{\partial^2 l}{\partial \mu_i \partial \mu_j} = \frac{1}{\sigma^2}, \]
\[ I_3 = \frac{\partial^2 l}{\partial \mu_j \partial \mu_i} = \frac{1}{\sigma^2}, \]
\[ I_4 = \frac{\partial^2 l}{\partial \mu_j \partial \sigma} = \frac{1}{\sigma^2}, \]
\[ I_5 = \frac{\partial^2 l}{\partial \sigma \partial \sigma} = \frac{1}{\sigma^2}. \]

where \( i, k = 1, \ldots, n \), present the gene and \( j = 1, \ldots, m \), shows the probe indicator, respectively. Moreover, \( I \) stands for the loglikelihood function \( lnL \). Then, the \( I \) matrix is derived via

\[ I = \begin{bmatrix}
\frac{\partial^2 l}{\partial \mu_i \partial \mu_j} & \frac{\partial^2 l}{\partial \mu_i \partial \sigma} & \frac{\partial^2 l}{\partial \mu_i \partial S_j} & \cdots & \frac{\partial^2 l}{\partial \mu_i \partial S_n} \\
\frac{\partial^2 l}{\partial \mu_j \partial \mu_i} & \frac{\partial^2 l}{\partial \mu_j \partial \sigma} & \frac{\partial^2 l}{\partial \mu_j \partial S_i} & \cdots & \frac{\partial^2 l}{\partial \mu_j \partial S_n} \\
\frac{\partial^2 l}{\partial \sigma \partial \mu_i} & \frac{\partial^2 l}{\partial \sigma \partial \mu_j} & \frac{\partial^2 l}{\partial \sigma \partial S_i} & \cdots & \frac{\partial^2 l}{\partial \sigma \partial S_n} \\
\frac{\partial^2 l}{\partial S_i \partial \mu_i} & \frac{\partial^2 l}{\partial S_i \partial \mu_j} & \frac{\partial^2 l}{\partial S_i \partial \sigma} & \cdots & \frac{\partial^2 l}{\partial S_i \partial S_n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 l}{\partial S_n \partial \mu_i} & \frac{\partial^2 l}{\partial S_n \partial \mu_j} & \frac{\partial^2 l}{\partial S_n \partial \sigma} & \cdots & \frac{\partial^2 l}{\partial S_n \partial S_n}
\end{bmatrix}. \]

In order to find the variance-covariance matrix, we partition the above matrix into the four sub-matrices. The details of this partition can be found in [18] and its application can be seen in [2, 3]. Once \( I \) is reorganized with respect to its sub-matrices, the variances of
the model parameters can be derived from simultaneous solutions of partial derivatives from the reorganized $I$. For an illustration, we show the variance of $\hat{\mu}_i$ as follows:

$$V(\hat{\mu}_i) = \frac{1}{C_0} \left[ \frac{-2v}{k\sigma} \sum_{i=1}^{n} \left( \frac{(MM_{(i)} - pS_i - \mu_i)^2}{k\sigma^2} \right) + \frac{2v}{k\sigma^2} \sum_{i=1}^{m} S_i \right]$$

where

$$C_0 = \left( \frac{\partial^2 l}{\partial \mu^2} \right)_{\mu = \mu_0} = \sum_{i=1}^{n} \left( \frac{(\partial^2 l/\partial \mu_i \partial \mu_i)}{-\partial^2 l/\partial S_i^2} \right)$$

In the estimation of the model parameters for each array, we check the possible shape parameters for the LTS distribution from 2 to 40 with the jump size 0.05. In the computation, we find that most of the array can be optimally defined under the shape parameter 40 or close to 40, which practically indicates the normal distribution.

As presented in Figure 1A, while the level of concentrations increases, a nonlinear pattern is seen in the structure of the estimated signals, like other methods. Figure 1B shows the associated graph via all other methods. Furthermore, the average estimated signals via multi-RGX versus nominal log-concentrations are plotted for each gene and the graph is reported in Figure 2.

Then to evaluate the performance in accuracy for each level of concentrations, we regress the spike-in genes which have the same concentration according to the distinct log-concentrations and apply the simple linear regression method. The computed $R^2$ is called as the signal detect $R^2$ and its associated slope term is named as the signal detect slope. Then, we compute the $R^2$ values and slope terms for each gene separately and later calculate the mean of both values individually. We call this measure as simply $R^2$. Hereby as seen in Table 1, the signal detect $R^2$, signal detect slope and $R^2$ are found as 0.79, 1.04 and 0.92, respectively. From the results, it is seen that $R^2$ is comparable with respect to its alternatives. Yet, under all concentrations simultaneously, the array indicates relatively more linear relation between estimated signals and concentrations on the nominal log-scale on the average, similar to multi-mgMOS, which has one of the perfection results for this type of modelling. But different from multi-mgMOS, our novel index is fully frequentist, i.e. based on the maximum likelihood approach. On the other

**Results**

**Application via a real dataset**

In the analysis, we use a benchmark Affymetrix spike-in dataset. This dataset consists of 11 spike-in genes coming from four bacterial ancestor genes and 59 arrays with 10,864 probe pairs. For the analysis, the 16 spike-in probe pairs are taken. Here, 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 [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 14 probes in each array, resulting in 224 observations per array.

Finally, for the analysis, we use the R programme language with our original codes for all calculations.

Hereby, in the estimation of the model parameters for each array, we check the possible shape parameters for the LTS distribution from 2 to 40 with the jump size 0.05. In the computation, we find that most of the array can be optimally defined under the shape parameter 40 or close to 40, which practically indicates the normal distribution.

As presented in Figure 1A, while the level of concentrations increases, a nonlinear pattern is seen in the structure of the estimated signals, like other methods. Figure 1B shows the associated graph via all other methods. Furthermore, the average estimated signals via multi-RGX versus nominal log-concentrations are plotted for each gene and the graph is reported in Figure 2.

Then to evaluate the performance in accuracy for each level of concentrations, we regress the spike-in genes which have the same concentration according to the distinct log-concentrations and apply the simple linear regression method. The computed $R^2$ is called as the signal detect $R^2$ and its associated slope term is named as the signal detect slope. Then, we compute the $R^2$ values and slope terms for each gene separately and later calculate the mean of both values individually. We call this measure as simply $R^2$. Hereby as seen in Table 1, the signal detect $R^2$, signal detect slope and $R^2$ are found as 0.79, 1.04 and 0.92, respectively. From the results, it is seen that $R^2$ is comparable with respect to its alternatives. Yet, under all concentrations simultaneously, the array indicates relatively more linear relation between estimated signals and concentrations on the nominal log-scale on the average, similar to multi-mgMOS, which has one of the perfection results for this type of modelling. But different from multi-mgMOS, our novel index is fully frequentist, i.e. based on the maximum likelihood approach. On the other
Figure 1: Average estimated spike-in signal versus concentrations on the nominal logarithmic scale (log2) for Affymetrix data via (A) multi-RGX and (B) MAS5.0, MBEI, RMA, GC-RMA, multi-mgMOS, FGX and RGX methods.

Figure 2: Signal versus concentrations on log-scale for Affymetrix data.

Table 1: Signal detect $R^2$, signal detect slope and average $R^2$, of all methods for Affymetrix data.

| Method         | Signal detect $R^2$ | Signal detect slope | $R^2$  |
|----------------|---------------------|---------------------|--------|
| Perfection value | 1.00                | 1.00                | 1.00   |
| MAS5.0         | 0.86                | 0.71                | 0.89   |
| RMA            | 0.80                | 0.63                | 0.99   |
| MBEI           | 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   |
| Multi-RGX      | 0.79                | 1.04                | 0.92   |

Table 2: Real computational time programme language of BGX, multi-mgMOS, FGX and multi-RGX, respectively, for Dataset Affymetrix data.

| Model          | Programme language | Computational time |
|----------------|--------------------|--------------------|
| BGX            | C++                | 32.5 h             |
| Multi-mgMOS    | R and C            | 50 min             |
| FGX            | R                  | 4 s                |
| Multi-RGX      | R                  | 34 s               |

hand, it indicates a stronger linearity between the relation of signals and concentrations with respect to the FGX and RGX outputs. This can be seen another advantage of multi-RGX regarding other choices in the analyses.

Besides, in order to assess the computational demand for alternative approaches, we compare the time for calculation via multi-RGX with other records in Table 2. It is seen that multi-RGX is significantly faster than BGX and multi-mgMOS. But with respect to the time for FGX, it is relatively slower. The reason is that we need to compute the optimal concomitant of probe sets, which converge to the true order in two or three iterations and this calculation causes extra time that is observed 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.

Application via simulated dataset

To assess the quality of our novel gene expression index, when the data are far from normal or have outliers, we simulate 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 every generated array. In the first data, we generate the PM and MM values
under the normal distribution and in the second data, we produce the values under the mean shifted normal density. Then, we mix these two sets in order to get two location-mixture datasets with different ratios of outliers. Hereby, the mathematical description of the first set is presented as below:

$$0.5N(S_i + \mu_H, \sigma^2) + 0.5N(S_i + \mu_R + \delta \sigma, \sigma^2), \quad (15)$$

and the second location mixture can be described as the following structure:

$$0.9N(S_i + \mu_H, \sigma^2) + 0.1N(S_i + \mu_R + \delta \sigma, \sigma^2), \quad (16)$$

where $N(\cdot,\cdot)$ indicates the univariate normal with the given parameters.

Here, the first data possess a large number of outliers. On the other hand, the second dataset owns a relatively moderate number of extreme observations. Moreover, in all simulations, we assume that every gene is measured under specific concentrations. Thereby, the data are simulated under $S_i$ setting to $S_i = 2, \ldots, 11$, for the gene $i = 1, \ldots, 10$, respectively, on the original scale for each presumed concentration level. Furthermore, $\mu_H, p, \sigma$ are equated as 1, 0.7 and 1, in order. Finally, the shift in the location, $\delta$, is set to $\delta = 10$ for both datasets.

In the assessment of accuracies for the estimated results, we select three criteria: the average error (AE), mean absolute error (MAE) and the root mean square error (RMSE) [19].

In Tables 3 and 4, we list the estimated values, i.e. $S_i$, $\mu_H, p$ and $\sigma$, of FGX, RGX and multi-RGX for each gene and model parameters, in order, for the expression in Equations (15)–(16), respectively, by computing the associated absolute errors.

### Table 3: Estimated model parameters of the first simulated location-mixture dataset (Expression in 15) via FGX, RGX and multi-RGX with their absolute errors (AE) and true values.

| Parameter | True value | FGX | AE of FGX | RGX | AE of RGX | Multi-RGX | AE of multi-RGX |
|-----------|------------|-----|-----------|-----|-----------|-----------|-----------------|
| $p$       | 0.7        | 0.88| 0.25      | 0.87| 0.24      | 0.88      | 0.25            |
| $\mu_H$   | 1          | -23.91| 24.91 | -23.91| 24.91 | -20.99| 21.99         |
| $\sigma$  | 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: Estimated model parameters of the second simulated location-mixture dataset (Expression in 16) via FGX, RGX and multi-RGX with their absolute errors (AE) and true values.

| Parameter | True value | FGX | AE of FGX | RGX | AE of RGX | Multi-RGX | AE of multi-RGX |
|-----------|------------|-----|-----------|-----|-----------|-----------|-----------------|
| $p$       | 0.7        | 0.95| 0.36      | 0.93| 0.33      | 0.95      | 0.36            |
| $\mu_H$   | 1          | -25.63| 26.63 | -25.63| 26.63 | -22.18| 23.18         |
| $\sigma$  | 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        |
From Tables 3 and 4, we observe that there is a large amount of bias in the estimates whose cause can be explained as follows: The first dataset is generated for 10 genes on the original scale, representing very low concentrations. Then, we take the nominal log-scale of all these simulated measurements in order to calculate the estimated signals. By this way, indeed, we evaluate the performance of all three models under very low concentrations since the estimates of all indices are problematic under these levels. Furthermore, as observed in Figure 1B, FGX, RGX and multi-RGX are able to measure the amount of the noisy signals on the array when there is no effect on the gene. Accordingly, from the outcomes we can see that this amount is approximately 21 units as measured under $\mu_H$, and if we subtract this shift from all estimated signals, we can find closer estimates of the signals with respect to their true values.

Additionally, in Table 5, we present the mean absolute error (MAE) and the root mean square error (RMSE) for the estimated signals via three selected models and in Table 6, 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, it is seen that the estimates of multi-RGX are more accurate than the ones computed via FGX and RGX. But the computational demand of multi-RGX is more than other two alternatives in the real time. However, this difference is indeed can be tolerable if we compare the results of the CPU time. This output shows that the real time of multi-RGX can be even improved with the help of an effective programming since the calculation of the estimates is almost performed under the same computer time in the end.

Additionally, to evaluate the performance of each alternative in a large dataset, we extend the calculation via the simulated data by 10,000 and 20,000 genes. Finally, we compute AE, MAE and RMSE for estimated $S_i$ in which the first data have 10,000 genes ($i=1, \ldots, 10,000$) and the second data have 20,000 genes ($i=1, \ldots, 20,000$). In contrast, for the remaining model parameters $p, \mu_H$ and $\sigma$, we compute directly AE, MAE and RMSE, which are unique per study.

Hereby, in the simulation we set the true value of $p$, $\mu_H$, and $\sigma$ to 0.7, 1 and 1, in order. For the true signals $S_i$, we initially generate values according to the number of total genes ranging from 2 to 11 from uniform distributions, assuming that the genes are measured under 10 distinct concentrations similar to the previous dataset. Also, the intensities are measured on the original scale due to the fact that we aim to evaluate the performance of all three methods under the most problematic range of intensities, i.e. the worst scenario for the comparison. Then, alike other estimates, 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 10,000 and 20,000 values for the first and second dataset, respectively. Whereas, they are directly computed for estimates of $p, \mu_H$ and $\sigma$. The results are presented in Tables 7–9 for the first large dataset (with 10,000 genes) whose mixture proportions are arranged according to Equations (15)–(16), in order. Moreover, the outcomes of the second large dataset (with 20,000 genes) are listed in Tables 9 and 10 whose mixture model is generated with indices.

| Table 5: Mean absolute error (MAE) and root mean square error (RMSE) of FGX, RGX and multi-RGX in the calculation of the estimated signals in two simulated location-mixture datasets. |
|---|---|---|
| Data | Model | 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 |

| Table 6: Real (in seconds) and central processing unit (CPU) time of FGX, RGX and multi-RGX in the calculation of the estimated model parameters in two simulated location-mixture datasets. |
|---|---|---|
| Data | Model | 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 7: Absolute error (AE) of FGX, RGX and multi-RGX in the calculation of the estimated $p$, $\mu$, and $\sigma$ in the large dataset with 10,000 genes according to two simulated location-mixture models.

| True value | FGX | AE of FGX | RGX | AE of RGX | Multi-RGX | AE of multi-RGX |
|------------|-----|-----------|-----|-----------|-----------|-----------------|
| $p$        | 0.7 | 1.14      | 0.63| 1.00      | 0.43      | 1.14            |
| $\mu$      | 1   | 5.64      | 4.65| 5.64      | 4.64      | 5.66            |
| $\sigma$   | 1   | 0.79      | 0.21| 0.24      | 0.76      | 0.23            |

Table 8: 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 $S_i$ ($i = 1, \ldots, 10,000$) in the large dataset with 10,000 genes according to the two simulated location-mixture models.

|             | FGX  | RGX  | Multi-RGX |
|-------------|------|------|-----------|
| First mixture | AAE  | MAE  | RMSE      |
| $S_1$      | 1.41 | 8.73 | 872.85    |
| $S_2$      | 1.31 | 8.16 | 816.43    |

Table 9: Absolute error (AE) of FGX, RGX and multi-RGX in the calculation of the estimated $p$, $\mu$, and $\sigma$ in the large dataset with 20,000 genes according to two simulated location-mixture models.

| True value | FGX | AE of FGX | RGX | AE of RGX | Multi-RGX | AE of multi-RGX |
|------------|-----|-----------|-----|-----------|-----------|-----------------|
| $p$        | 0.7 | 1.14      | 0.63| 1.00      | 0.43      | 1.14            |
| $\mu$      | 1   | 5.64      | 4.64| 5.64      | 4.64      | 5.64            |
| $\sigma$   | 1   | 0.78      | 0.22| 0.24      | 0.77      | 0.24            |

Table 10: 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 $S_i$ ($i = 1, \ldots, 20,000$) in the large dataset with 10,000 genes according to the two simulated location-mixture models.

|             | FGX  | RGX  | Multi-RGX |
|-------------|------|------|-----------|
| First mixture | AAE  | MAE  | RMSE      |
| AAE         | 1.41 | 1.41 | 1.45      |
| MAE         | 8.69 | 8.69 | 8.70      |
| RMSE        | 1229.42 | 1129.42 | 1230.33   |
| Second mixture | AAE  | MAE  | RMSE      |
| AAE         | 1.32 | 1.32 | 1.34      |
| MAE         | 8.20 | 8.20 | 8.20      |
| RMSE        | 1159.14 | 1159.14 | 1160.36   |

Conclusion

In this study, we have proposed a new gene expression index for one-channel microarrays, called multi-RGX, which uses the long-tailed symmetric distribution to describe the signals and takes into account both gene and probe specific measurements. Moreover, multi-RGX generates an explicit expression for each model parameter.

In the assessment of our proposal gene expression, we have compared the accuracy and the computational demand under certain criteria. According to the results, we have observed that multi-RGX can keep a high accuracy regarding alternative gene expression indices while preserving less computational cost under small datasets.
But there is no significant difference among any model in large data from the simulation analyses.

As the extension of this study, we consider to implement the multiple adaptive regression splines (MARS) approach to model the observed nonlinear relationship between signals and concentrations, in particular, under low and high concentrations. This idea has been previously performed in the study of Kunjdaje et al. [20] to detect the function of histone modification levels. We believe that similar improvement in the extension and the proposed current model can open new avenues in the analyses of microarray for different biological researches.

**Conflict of interest statement:** There are no conflicts of interest among the authors.

**References**

1. Sanchez A, Ruiz de Villa MC. A tutorial review of microarray data analysis. Technical report, Universitat de Barcelona, 2008.
2. Purutçuoğlu V, Wit E. FGX: a frequentist gene expression index for Affymetrix arrays. Biostatistics 2007;8:433–7.
3. Purutçuoğlu V. Robust gene expression index. Math Probl Eng 2012;2012:1–17.
4. Li C, Wong W. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 2001;98:31–6.
5. Hubbell E, Liu W-M, Mui R. Robust estimators for expression analysis. Bioinformatics 2002;18:1585–92.
6. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;4:249–64.
7. Milo M, Fazeli A, Niranjan M, Lawrence ND. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. Biochem Soc Trans 2003;31:1510–2.
8. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A model-based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc 2004;99:909–17.
9. Liu X, Milo M, Lawrence N, Rattray M. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. Bioinformatics 2005;21:3637–44.
10. Hein AM, Richardson S, Causton HC, Ambler GK, Green PJ. BGX: a fully Bayesian gene expression index for Affymetrix GeneChip data. Biostatistics 2005;6:349–73.
11. Binder H, Preibisch S. “Hook”-calibration of GeneChip-microarrays: theory and algorithm. Algorithms Mol Biol 2008;3:12.
12. Kohl M, Deigner HP. Preprocessing of gene expression data by optimally robust estimators. BMC Bioinformatics 2010;11:1–15.
13. Lahti L, Torrente A, Elo LL, Brazma A, Rung J. A fully scalable online pre-processing algorithm for short oligonucleotide microarray atlases. Nuc Acids Res 2013;41:e110.
14. Akal T. Gene expression indices for single-channel microarrays. Master’s thesis, Middle East Technical University, 2013.
15. Tiku ML, Tan WY, Balakrishnan N. Robust inference. New York: Marcel Dekker, 1986.
16. Bhattacharyya GK. The asymptotics of maximum likelihood and related estimators based on type II censored data. J Am Stat Assoc 1985;80:398–404.
17. Bain LJ, Engelhardt M. Introduction to probability and mathematical statistics. CA: Duxbury Classic Series, 1991.
18. Healy MJ. Matrices for statistics. New York, USA: Oxford University Press, 2000.
19. Batmaz İ, Kartal-Koç E, Weber GW. Advances in Intelligent Modelling and Simulation: Studies in Computational Intelligence, chapter Robust Regression Metamodelling of Complex Systems: The Case of Solid Rocket Motor Performance Metamodelling. Berlin Heidelberg: Springer-Verlag, 2012:221–251.
20. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol 2012;13:1–17.