Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2012, Article ID 182758, 12 pages
doi:10.1155/2012/182758

Research Article

Robust Gene Expression Index

Vilda Purutçuoğlu

Department of Statistics, Middle East Technical University, 06531 Ankara, Turkey

Correspondence should be addressed to Vilda Purutçuoğlu, vpurutcu@metu.edu.tr

Received 10 October 2011; Accepted 22 October 2011

Academic Editor: Gerhard-Wilhelm Weber

The frequentist gene expression index (FGX) was recently developed to measure expression on Affymetrix oligonucleotide DNA arrays. In this study, we extend FGX to cover nonnormal log expressions, specifically long-tailed symmetric densities and call our new index as robust gene expression index (RGX). In estimation, we implement the modified maximum likelihood method to unravel the elusive solutions of likelihood equations and utilize the Fisher information matrix for covariance terms. From the analysis via the bench-mark datasets and simulated data, it is shown that RGX has promising results and mostly outperforms FGX in terms of relative efficiency of the estimated signals, in particular, when the data are nonnormal.

1. Introduction

Microarray technology enables the measurement of RNA (transcripted DNA) expression levels. For this purpose, it uses different kinds of optical techniques, which quantify the colour intensities on the array. These intensities can be used to capture the functional homogenous subgroups of genes via various clustering algorithms [1, 2] and to model the uncertainty in the associated gene networks with the help of different optimization techniques [3, 4]. But because of the distinct experimental conditions, those measured intensities include different sources of errors, some of which are random and some of which are systematic. The former errors do not change the overall mean accuracy of the results and cannot be removed from the measurements. On the contrary, the latter causes a systematic bias if included [5]. Fortunately, they can be eliminated through methods of normalization. The Affymetrix GeneChip is the most common oligonucleotide array, where each array is composed of small strings of DNA, each 25 base pairs long that bind to complementary transcripts, thereby measuring transcription from DNA to RNA for each gene. Each gene on the array is represented by 11 to 20 probe pairs. Each pair consists of a perfect match (PM) and a mismatch (MM) probe. The PM is designed to measure the amount of gene transcription

plus some additional nonspecific binding. The MM probe, which only differs from its PM probes by the 13th base pair, is designed to measure the amount of nuisance or background signal. But it has been recognized [6, 7] that the MM values are heavily correlated with the PM values, suggesting that they also contain a fraction of the original true gene expression signal. In order to describe the true gene expression level by modelling the probe effect in each array, and calculating the intensities in terms of PM and MM, statistics are needed. This common statistics is called the gene expression index. There are a number of methods, also called models, to summarize the multiple probe pair information into a single gene expression value, that is, a gene index. MAS 5.0 [8, 9] is one of the most common methods which assume true signals in the PM probes corrupted in an additive way by background signals which are merely measured in the MM values. If the intensities are negatives, that is, MM > PM, the methods suggests a background substraction from the PM values. RMA (robust microarray analysis) [10] is the first method which uses no substraction from PM values when MM > PM, whereas it considers that PM is the only source of true signals and the MM values as a measure of background signal is dubious, thereby should be ignored. GC-RMA (robust microarray analysis based on GC content) [10] is the first method which takes into account the existence of true signal in MM intensities. In this method, PM's are found by the summation of optical noise, nonspecific hybridization, and the true signal. But MM's are also accepted to have a fraction of the true signal under the assumption of log-normal distribution for both MM's and PM's. Later BGX (Bayesian gene expression index) [6] and multi-mgMOS (multiple array mgMOS) [7] models use the same idea for estimation. BGX describes PM and MM via truncated normal density on the logarithmic scale by guaranteeing the nonnegativity of true signals and nonspecific hybridization, whereas multi-mgMOS considers gamma distributed intensities on the original scale. Hereby, the main difference between BGX and multi-mgMOS is their way of inference for the model parameters in the sense that BGX implements a fully Bayesian approach for the estimation, thereby faces with the challenge of computational demand, and multi-mgMOS performs the maximum a posterior probability (MAP) which enables us to use less computational cost with respect to the BGX calculation. In the FGX model, by using the same idea for the description of intensities, it is assumed that the log-expressed intensities are normal as

$$\log \mathrm{PM}_{ij} \sim N\left(S_i + \mu_H, \sigma^2\right), \qquad \log \mathrm{MM}_{ij} \sim N\left(pS_i + \mu_H, \sigma^2\right), \tag{1.1}$$

where S_i represents the true expression value for the ith gene, p stands for the fraction of the specific hybridization to the MM probe, and μ_H is the mean of the nonspecific hybridization, which shows different sources of nuisance intensities. i and j display the gene indicator ($i = 1, \ldots, n$) and the probe indicator ($j = 1, \ldots, m$), respectively. Finally, σ^2 denotes the model variances of normally distributed PM and MM intensities [11]. When observing probe summaries as their means, rather than their individual probe values, we could adjust (1.1) by replacing σ^2 by σ^2/m, whereby m is the number of probes in the probe set. In this study, we extend this model by relaxing the normality assumption. We allow the logarithms of PM and MM to have long-tailed symmetric (LTS) densities, thereby covering distributions ranging from normal to cauchy. In inference of the model parameters, we implement the modified maximum likelihood estimators (MMLE) [12] due to the fact that the likelihood equations under LTS density do not have explicit solutions. We evaluate the performance of our model in benchmark spike-in and simulated datasets. From the analysis we conclude that the RGX

model is promising in terms of accuracy and can be a helpful tool for the biomolecular engineering's application in computational biology and bioinformatics.

2. Robust Gene Expression Index

In order to estimate the gene expression level of a transcript from perfect PM (perfect matches) and MM (mismatches) values, typically, it is suggested that the intensities are distributed via gamma [7, 13] on the original scale or normal [4, 6, 8] on the logarithmic scale. On the other side, the models suggested by [8, 10] do not use any distributional assumption for modeling the intensities. On the contrary, they implement robust estimators or some optimization techniques to find the true gene expressions. However, from the study of [6] whose inference is computed by the MCMC (Markov chain Monte Carlo) algorithm, it is suggested that the true distribution of the intensities can be originated from the truncated normal, and in comparison to MAS 5.0 (Microarray Suite Software), MBEI (model-based gene expression index) [14], and RMA (robust microarray analysis), the point estimates of the posterior distributions of gene expression indices via BGX (Bayesian gene expression index) perform better, in particular, to detect the differences at low levels. Moreover, both BGX and RMA give biggest differences when the genes are ranked according to the degree of differential expression for every possible pairwise comparison of genes. This finding is interesting in the sense that the models which do not depend on the strict normality assumption outperform in comparison with other indices. In this study, to decide on the distribution of intensities on the logarithmic scale ($\log_2$), we consider to draw the quantile-quantile (Q-Q) plot of the data and compare it with the normal density line. From the results, it is seen that PM and MM of Affymetrix probes deviate from the straight line mostly at the tails, which is the property of LTS (long-tailed symmetric) distribution. Hereby, we model the intensities as shown in (2.1) and call it the robust gene expression index (RGX), as we consider both normal and its plausible alternatives in inference of the true signals. In this way, we get resistent estimates for departures from normality. In (2.1), similar to (1.1), S_i and μ_H describe the true signal for the ith gene and nonspecific hybridization, respectively. Moreover, p indicates the fraction of the true signal in MM probes, and σ^2 denotes the variances of both PM's and MM's:

$$\log \mathrm{PM}_{ij} \sim \mathrm{LTS}\left(S_i + \mu_H, \sigma^2\right), \qquad \log \mathrm{MM}_{ij} \sim \mathrm{LTS}\left(pS_i + \mu_H, \sigma^2\right). \tag{2.1}$$

2.1. Estimation via MMLE Method

In order to infer the model parameters, we summarize the probe values by taking their means like FGX (frequentist gene expression index) seeing that the typical analysis of the Affymetrix data is conducted on a probe set, rather than an individual probe level. Then, we define the likelihood function L below conditional on perfect matches $\mathrm{PM} = (\mathrm{PM}_1, \ldots, \mathrm{PM}_n)$ and mismatches $\mathrm{MM} = (\mathrm{MM}_1, \ldots, \mathrm{MM}_n)$ for each array, where $\mathrm{PM}_i := \sum_{j=1}^{m} \mathrm{PM}_{ij}/m$, $\mathrm{MM}_i := \sum_{j=1}^{m} \mathrm{MM}_{ij}/m$, and $i = 1, \ldots, n$. In (2.2), we assume that the expression of every gene in an oligonucleotide is independent on each other similar to BGX (Bayesian gene expression index), mgMOS (modified gamma model for oligonucleotide signal), and multi-mgMOS

(multiple array mgMOS) models. But unlike these indices, our index computes single array, rather than multiple arrays simultaneously, at a time,

$$L(\underline{S}, \mu_H, p, \sigma|\mathrm{PM}, \mathrm{MM}) \propto \left(\frac{\sqrt{m}}{\sigma}\right)^n \prod_{i=1}^{n}\left(1+\frac{z_{\mathrm{PM}_i}^2}{k}\right)^{-v} \times \left(\frac{\sqrt{m}}{\sigma}\right)^n \prod_{i=1}^{n}\left(1+\frac{z_{\mathrm{MM}_i}^2}{k}\right)^{-v}, \tag{2.2}$$

in which v shows the shape parameter ($v \geq 2$) assuring the existence of μ and $k = 2v - 3$. $\underline{S} = (S_1, \ldots, S_n)$ is the n-dimensional vector of the true signals. $z_{\mathrm{PM}_i} = (\mathrm{PM}_i - S_i - \mu_H)/(\sigma/\sqrt{m})$ and $z_{\mathrm{MM}_i} = (\mathrm{MM}_i - pS_i - \mu_H)/(\sigma/\sqrt{m})$ represent the standardized values of PM and MM intensities for $i = 1, \ldots, n$, respectively. In inference of the unknown parameters μ_H, p, S_i ($i = 1, \ldots, n$), and σ, we derive the following partial loglikelihoods:

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu_H} &= \frac{2v\sqrt{m}}{\sigma k}\sum_{i=1}^{n}\left[g(z_{\mathrm{PM}_i}) + g(z_{\mathrm{MM}_i})\right], \\ \frac{\partial \ln L}{\partial p} &= \frac{2v\sqrt{m}}{\sigma k}\sum_{i=1}^{n} S_i\left[g(z_{\mathrm{MM}_i})\right], \end{aligned} \tag{2.3}$$

where $g(z_{\mathrm{PM}_i}) = z_{\mathrm{PM}_i}/\{1 + (1/k)z_{\mathrm{PM}_i}^2\}$ and $g(z_{\mathrm{MM}_i}) = z_{\mathrm{MM}_i}/\{1 + (1/k)z_{\mathrm{MM}_i}^2\}$. When we equate these expressions to zero in order to find the maximum likelihood estimates of the model parameters, it is seen that the loglikelihood derivations do not have explicit solutions and the iterative methods are needed for approximately solving the equations. In this study, we overcome the underlying challenge by the MML (modified maximum likelihood) method which is asymptotically equivalent to the maximum likelihood estimates [15, 16]. Moreover, for small sample size, this method produces estimates as efficient as the maximum likelihood results. In the MML technique, briefly, we use the ordered variate of residuals $e_{\mathrm{PM}_{(i)}} = \mathrm{PM}_{[i]} - S_{[i]} - \mu_H$ and $e_{\mathrm{MM}_{(i)}} = \mathrm{MM}_{[i]} - pS_{[i]} - \mu_H$ by replacing z_{PM_i} by $z_{\mathrm{PM}_{[i]}} = (\mathrm{PM}_{[i]} - S_{[i]} - \mu_H)/(\sigma/\sqrt{m})$ and z_{MM_i} by $z_{\mathrm{MM}_{[i]}} = (\mathrm{MM}_{[i]} - pS_{[i]} - \mu_H)/(\sigma/\sqrt{m})$, respectively. In these expressions $(\mathrm{PM}_{[i]}, \mathrm{MM}_{[i]}, S_{[i]})$ are the concomitant observations of the corresponding ith ordered (in increasing magnitude) $e_{(i)}$'s. The method takes the linear approximation of the $g(z_{\mathrm{PM}_i})$ and $g(z_{\mathrm{MM}_i})$ functions by the first-order Taylor expansion around the ith population quantile $t_{(i)}$ of the Student's t-distribution with $(2v - 1)$ degrees of freedom. Hereby, the nonlinear functions are approximated by

$$g\left(z_{\mathrm{PM}_{[i]}}\right) \simeq \alpha_i + \beta_i z_{\mathrm{PM}_{[i]}}, \qquad g\left(z_{\mathrm{MM}_{[i]}}\right) \simeq \alpha_i + \beta_i z_{\mathrm{MM}_{[i]}}, \tag{2.4}$$

where

$$\alpha_i = \frac{2t_{(i)}^3/k}{\left(1 + t_{(i)}^2/k\right)^2}, \qquad \beta_i = \frac{1 - t_{(i)}^2/k}{\left(1 + t_{(i)}^2/k\right)^2}, \tag{2.5}$$

$\sum_{i=1}^{n} \alpha_i = 0$ because of symmetry. Accordingly, the closed form of μ_H is found as $\widehat{\mu}_H = (\sum_{i=1}^{n} \beta_i \mathrm{MM}_{[i]} - \widehat{p} \sum_{i=1}^{n} \beta_i \mathrm{PM}_{[i]})/((1 - \widehat{p}) \sum_{i=1}^{n} \beta_i)$. On the other side in the estimation of σ, the

mean probes, that is, PM_i and MM_i, are not sufficient statistics. We regain the lost information in inference of σ, by recomputing its MML derivation via the complete loglikelihood. Then, we express the partial derivative of σ in terms of ordered variates j via $z_{\mathrm{PM}_{i[j]}} = (\mathrm{PM}_{i[j]} - S_i - \mu_H)/\sigma$ and $z_{\mathrm{MM}_{i[j]}} = (\mathrm{MM}_{i[j]} - pS_i - \mu_H)/\sigma$. In the end, we get $\widehat{\sigma} = (B + \sqrt{B^2 + 4nmC})/(2nm)$, where $C = (v/k)\{\sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(\mathrm{PM}_{i[j]} - \widehat{S}_i - \widehat{\mu}_H)^2 + \sum_{i=1}^{n}\sum_{j=1}^{m}\beta_j(\mathrm{MM}_{i[j]} - \widehat{p}\widehat{S}_i - \widehat{\mu}_H)^2\}$ and $B = (v/k)\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j(\mathrm{PM}_{i[j]} - \mathrm{MM}_{i[j]})$.

Finally, for the inference of S_i, we solve the partial derivative of loglikelihood with respect to S_i by taking the sufficient statistics of σ. So, the estimate of S_i is described as

$$\widehat{S}_i = \frac{\widehat{\sigma}(1+\widehat{p})\alpha_i + (\mathrm{PM}_{[i]} + \widehat{p}\mathrm{MM}_{[i]})\beta_i - \widehat{\mu}_H(1+\widehat{p})\beta_i}{(1+\widehat{p}^2)\beta_i}. \tag{2.6}$$

To infer $\widehat{p}$, we follow a two-stage procedure. In the first stage, we give initial values for $\widehat{\mu}_H$, $\widehat{\sigma}$, $\widehat{S}_i$, $\widehat{p}$ which are selected as their estimates under normality and find the candidate values of α's, β's, and true concomitants used in the MML estimation. Then, we compute the MML estimates of $\widehat{\mu}_H$, $\widehat{\sigma}$, and $\widehat{S}_i$, by taking previous estimates of α's, β's, and concomitants as the initial values for the next iteration of the first stage. This procedure is repeated until both concomitants and MML estimates are stabilized. From the findings, we observe that, in general, three iterations are enough to get stable results. In the second stage, final MML estimates of μ_H, σ, and S_i from the first step are used in $\partial \ln L/\partial p$. On the other hand, the true p is the one which maximizes this expression within $0 \le p \le 1$ with a step size 0.001, thereby $\widehat{p}$ that gives the closest value to zero in $\partial \ln L/\partial p$ is taken as the MML estimate of p.

2.2. Observed Fisher Information Matrix

The MML (modified maximum likelihood) estimators are asymptotically equivalent to the ML (maximum likelihood) estimators [12, 16], resulting in the maintenance of the minimum variance bound and unbiasness properties. Due to its full efficiency, the covariances and variances of the estimators can be found via the inverse of the Fisher information matrix I, I^{-1}. Whereas since we have a finite number of samples, we implement the observed I

$$I = -\begin{bmatrix} \dfrac{\partial^2 l}{\partial \mu_H^2} & \dfrac{\partial^2 l}{\partial \mu_H \partial p} & \dfrac{\partial^2 l}{\partial \mu_H \partial S_1} & \dfrac{\partial^2 l}{\partial \mu_H \partial S_2} & \cdots & \dfrac{\partial^2 l}{\partial \mu_H \partial S_n} \\ \dfrac{\partial^2 l}{\partial p \partial \mu_H} & \dfrac{\partial^2 l}{\partial p^2} & \dfrac{\partial^2 l}{\partial p \partial S_1} & \dfrac{\partial^2 l}{\partial p \partial S_2} & \cdots & \dfrac{\partial^2 l}{\partial p \partial S_n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \dfrac{\partial^2 l}{\partial S_n \partial \mu_H} & \dfrac{\partial^2 l}{\partial S_n \partial p} & 0 & 0 & \cdots & \dfrac{\partial^2 l}{\partial S_n^2} \end{bmatrix}_{(n+2)\times(n+2)}. \tag{2.7}$$

In (2.8), the variance of $\widehat{\mu}_H$ is given as an example. In this expression, C_0 shows a common constant term in all variances and covariances. $T_0 = (\mathrm{MM}_i - \widehat{p}\widehat{S}_i - \widehat{\mu}_H)/(k\widehat{\sigma})$,

$T_1 = (\mathrm{MM}_i - \hat{p}\widehat{S}_i - \hat{\mu}_H)^2/(k\hat{\sigma}^2)$, and $T_2 = (\mathrm{PM}_i - \widehat{S}_i - \hat{\mu}_H)^2(k\hat{\sigma}^2)$, where PM and MM denote the perfect matches and mismatches probes, respectively, as previously used

$$V(\hat{\mu}_H) = \frac{1}{C_0}\left[\frac{2v}{k\hat{\sigma}^2}\sum_{i=1}^{n}\widehat{S}_i^2\frac{1-T_1}{(1+T_1)^2} - \sum_{i=1}^{n}\frac{-(2v/k\hat{\sigma})(T_0/(1+T_1)) + (2v\hat{p}/k\hat{\sigma}^2)\widehat{S}_i\left((1-T_1)/(1+T_1)^2\right)}{(2v/k\hat{\sigma}^2)\left((1-T_2)/(1+T_2)^2\right) + (2v\hat{p}^2/k\hat{\sigma}^2)\left((1-T_1)/(1+T_1)^2\right)}\right]. \tag{2.8}$$

2.3. Data Description in the Application

In the assessment of the MML (modified maximum likellihood) estimators, we use three datasets. The first two data are chosen by the other methods for the comparison, and the third data are generated by simulation and are used for the comparison between FGX (frequentist gene expression index) and RGX (robust gene expression index). In the first analysis, we implement a bench-mark Affymetrix spike-in data which have 59 arrays with 10864 probe sets. The data are available from http://affycomp.biostat.jhsph.edu/. For the evaluation, we use the common 16 spike-in probesets (numbered as 3777, 684, 1597, 38734, 39058, 36311, 36889, 1024, 36202, 36085, 40322, 407, 1091, 1708, 33818, and 546) whose concentration levels are publicly available. These spike-in genes are measured under 14 concentration levels listed as 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). Every gene is described by 16 probes in each array. In the second analysis, we use a GeneLogic spike-in dataset which has 14 arrays (arrays 92453, 92454, 92456, 92458, 92460, 92462, 92464, 92466, and 92491–92496 with 9 suffix hgu95a11) with 11 GeneLogic spike-in probes sets (viz. BioB-5, BioB-M, BioB-3, BioC-5, BioC-5, BioC-3, BioDn-3, DapX-5, DapX-M, DapX-3, CreX-5, and CreX-3 with affix AFFX-) whose concentration levels are publicly available and used for the evaluation of other methods [17]. In this dataset, except CreX-3 probe set, every spike-in gene is hybridized at 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 is composed of 20 probes. In the assessment, similar to other findings from different indices [6], the array 92466 and the spike-in gene DapX-M are excluded. Finally, in the third analysis, we use a simulated dataset which is a location mixture of two normal distributions with $0.5N(S_i + \mu_H, \sigma^2) + 0.5N(S_i + \mu_H + \delta\sigma, \sigma^2)$ structure for perfect matches PM and mismatched MM values. Here, δ stands for the constant affecting the location. In this set, we take 10 genes where each gene has 20 probes and S_i is accepted as $S_1 = 2, 3, \ldots, 13$ for $i = 1, 2, \ldots, 10$, respectively, assuming that every gene gives intensities under a specific concentration. Then, we set other unknown parameters to $\mu_H = 1$, $\sigma^2 = 1$, $p = 0.7$, and $\delta = 10$ considering that the second part of the mixture causes extreme observations with probability 0.5.

2.4. Assessment Criteria in Application

In order to evaluate our results by using the first Affymetrix dataset, we compare RGX (robust gene expression index) estimates with MAS 5.0 (Microarray Suite Software), MBEI (model-based gene expression index) or dChip, RMA (robust microarray analysis), GC-RMA (robust

mocroarray analysis on GC content), mgMOS (modified gamma model for oligonucleotide signal) [7], and multi-mgMOS (multiple array mgMOS) [7] results. For the comparisons of the first dataset via all these well-known methods, we use the following criteria which are presented in the table http://affycomp.biostat.jhsph.edu/AFFY2/TABLES/0.html and in the study of [18]: (i) signal detect that is found by regressing the gene expressions of all arrays on their corresponding nominal log concentrations, (ii) signal detect slope which is the slope term computed from this regression, (iii) R^2 that is found by taking the average of derived from each array separately, and (iv) low slope that is the slope term obtained as described above but for the genes under low concentrations ($0.25 \leq x \leq 16.0\,\text{pM}$). For the assessment, furthermore, we use three plots. In the first plot, we draw the average intensities of 14 Affymetrix spike-in probesets versus nominal log concentrations. This figure (Figure 1(a)) corresponds to the regression line which gives the signal detect R^2 in item. In the second plot, we draw the observed fold change across nominal fold change, where the genes are exposed in the same number of concentrations after the cancelation of zero concentrations on the original scale (Figures 1(b) and 1(c)). This type of the plot is used to get a prior information to find the most interesting genes which give the highest fold changes [18]. Finally, in the third plot, we compare the sensitivity of all methods by the average receiving operating characteristic (ROC) curve. For the analysis, we take the absolute difference of the same gene intensities in the two different arrays i and j. These differences are computed for all possible pairs ($i < j$) and ordered in increasing magnitude. Then, the number of true positives along every possible value of false positive from 0 to 100 is calculated. This process is implemented for each pair of arrays, and the average of true positives across every false positive value is plotted. In the analysis of the second dataset, we initially compare the computational time of BGX and multi-mgMOS with the results of FGX (frequentist gene expression index) and RGX (robust gene expression index). Then, to assess the relation between signals and concentrations, we draw the plot of the average estimated signals per concentration and compute the associated slope term and R^2. This comparison is based on the results of BGX (Bayesian gene expression index) presented in [6] and FGX given in [11]. To evaluate the performance of every gene with their variances, we plot the graph indicating the estimated intensities within a 95% confidence interval. Finally, in order to compare merely RGX and FGX when the data become far from normality or have outliers, we evaluate the simulated dataset. For the assessment, we repeat the simulation 10,000 Monte Carlo times and calculate the mean and standard deviation of the estimated model parameters. The results are compared with the associated true values in terms of accuracy, efficiency, and relative efficiency (RE). In the calculation of RE, we use RE = 100 (Variance of RGX/Variance of FGX).

3. Results

In RGX (robust gene expression index), since p and σ are the common parameters for the data, they might be affected by which probesets are included in inference. In our evaluation, we compute these common terms by using the selected spike-in genes in each dataset. We obtain the true v from the likelihood function of the long-tailed symmetric density. In order to find the best choice for v in which $v = \infty$ refers to the normal density, we calculate the loglikelihood, $\ln L$, score for every value of v from 2 to 52 with a step size 0.5 by setting the model parameters in $\ln L$ to their FGX (frequentist gene expression index) estimates. Accordingly, the true v can be the value which maximizes $(1/n)\ln L$ seeing that the highest likelihood information can be gathered under the most plausible $\widehat{v}$. From this searching

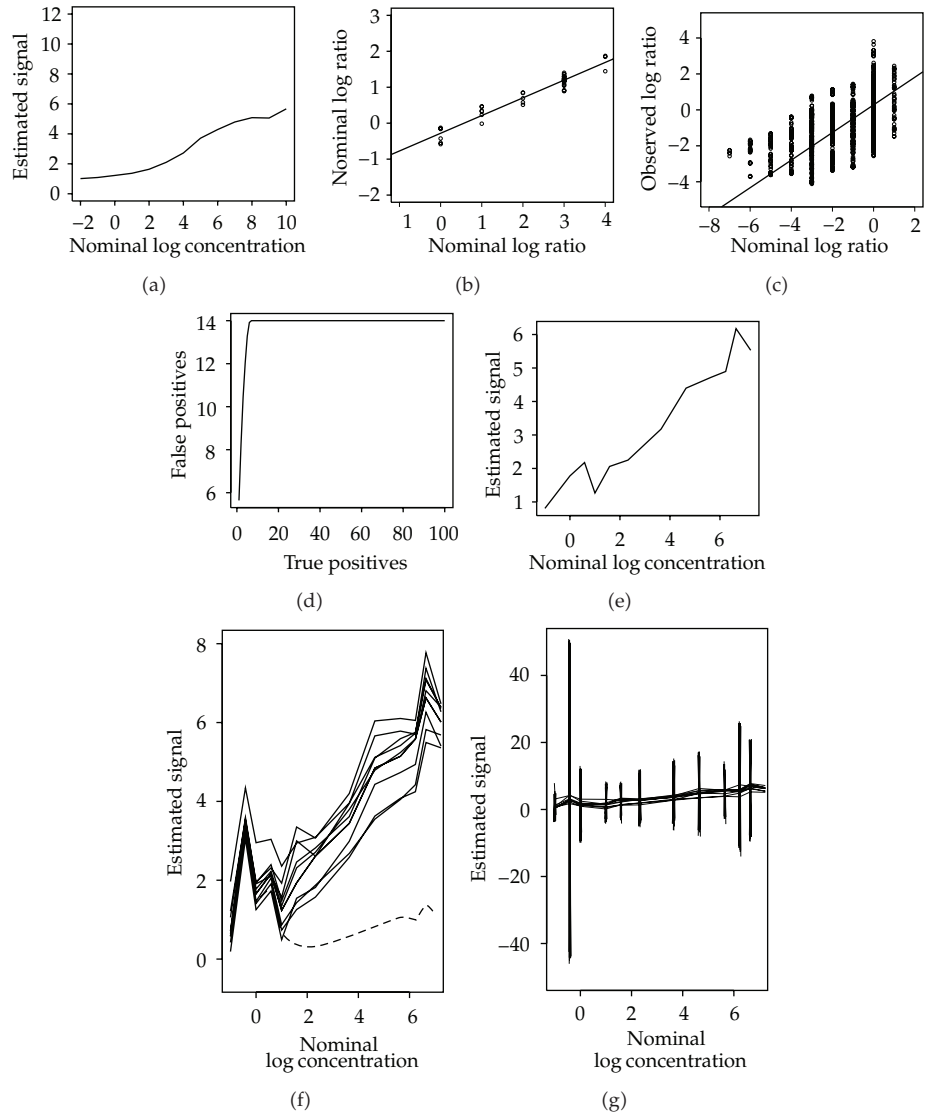

Figure 1: (a) Average estimated intensities of Affymetrix genes (except 3818 and 546). (b) Observed fold changes versus fold changes and fitted simple linear regression models for genes 684 and 1597. (c) Genes 38734, 39058, 36311, 36889, 1024, 36202, 36085, 40322, and 1708 in 59 Affymetrix arrays. (d) Average ROC curve. (e) GeneLogic data for RGX average estimated signals per nominal concentration. (f) Weighted average intensities of genes. (g) 95% confidence intervals.

process for both Affymetrix and GeneLogic datasets, we see that $v = 52$ is the optimal preference for the true v for all arrays. From the assessment of the first data, we observe that only FGX and RGX measure the zero signal among MAS 5.0 (Microarray Suite Software), RMA (robust microarray analysis), MBEI (model-based gene expression index), GC-RMA (robust microarray analysis based on GC content), mgMOS (modified gamma model for oligonucleotide signal), and multi-mgMOS (multiple array mgMOS) methods when the concentrations are negligibly small. Because the structure of both models enables us to

Table 1: Selected criteria with perfection values in Section 2.4 for Affymetrix data.

Method	Signal detect R^2	Signal detect slope	R^2	Low slope
MAS 5.0	0.86	0.71	0.89	0.72
RMA	0.80	0.63	0.99	0.29
MBEI (dChip)	0.85	0.53	0.99	0.25
GC-RMA	0.84	0.97	0.99	0.73
mgMOS	0.82	0.76	0.96	0.77
multi-mgMOS	0.80	1.03	0.96	1.21
FGX	0.94	0.43	0.90	0.26
RGX	0.96	0.44	0.92	0.27

compute the average value of μ_H under every concentration. Apart from the intercept term, we find a high similarity between all models [6]. With respect to the plot in Figures 1(b) and 1(c), we observe a straight line indicating a fitted simple linear regression line according to the given changes. From the selected criteria for the Affymetrix data, our results together with its strong alternatives are presented in Table 1, where the signal detect R^2 of RGX is better than all other alternatives. Furthermore, its average R^2 has comparable value and improves the results of FGX. If the estimates under low, medium, and high intensities are checked separately, it is seen that R^2 of each group is high in the sense that R^2 of medium (0.98) and high (1) intensities indicate almost perfect correlation and R^2 of low (0.88) intensities is relatively small.

Also, from the slope terms, we find that the relation between signals and concentrations is not linear on both original and nominal log scale. Finally, from the plots of the average ROC (receiving operating characteristic) curve (Figure 1(d)), the sensitivity of RGX is as good as FGX and RMA models. On the other hand, in the analysis of the GeneLogic data, we evaluate the computational time, and we find that both FGX (1 sec in R) and RGX (6 sec in R) are much faster than BGX (Bayesian gene expression index) (70 min in C++) and multi-msMOS (3 min in R). Then, we assess the plot of the average estimated signals per concentration. The resulting plot (Figure 1(e)) has $R^2 = 0.94$ with the slope term 0.62 implying a nonlinear relationship between signals and concentration, similar to the analysis via the Affymetrix spike-in data. In terms of the slope, RGX is slightly better than BGX (around 0.50) and FGX (around 0.60) [10]. Whereas apart from the estimation under low concentrations, we observe that the signals display a linear relation across concentrations. Finally, to evaluate the estimated intensities within a 95% confidence interval, we present Figure 1(f). In the computation of the variance, we give a weight in each gene in the sense that every estimated signal is weighted by the precision of all other signals at the same concentration. The analysis shows that although the estimates under low concentrations are affected by noise, resulting in larger confidence intervals, the estimates from the medium concentrations are precise and the ones from high concentrations are relatively better than the estimates under low concentrations. But we observe that both FGX and RGX indices have close performance [11]. Similar to the Affymetrix analysis, we anticipate this result. Because in both datasets, we see that the intensities indicate high v values, meaning that they are close to the normal density. Whereas in order to compare the performance of RGX and FGX when the data become far from normality, we use the simulated dataset whose model parameters are evaluated based on the mean, standard deviation, and relative efficiency from 10,000 Monte Carlo runs. In Table 2, we display that RGX and FGX have very close accuracies whereas RGX outperforms FGX in terms of efficiency when the number of extreme observation increases. The gain in

Table 2: Mean, standard deviation (Std. dev.), and relative efficiency (RE) of RGX and FGX estimates.

		RGX		FGX		
Parameter	True value	Mean	Std. dev.	Mean	Std. dev.	RE
p	0.7	0.735	0.042	0.735	0.042	100.000
μ_H	1	4.940	1.308	4.941	1.307	100.153
σ	1	4.972	0.073	3.847	484	2.275
S_1	2	3.124	1.529	3.146	1.551	97.183
S_2	3	4.116	1.574	4.133	1.597	97.140
S_3	4	5.080	1.637	5.090	1.663	96.898
S_4	5	6.085	1.648	6.092	1.670	97.383
S_5	6	7.060	1.687	7.061	1.709	97.442
S_6	7	8.054	1.765	8.050	1.788	97.444
S_7	8	9.045	1.780	9.036	1.802	97.573
S_8	9	10.029	1.819	10.015	1.840	97.730
S_9	10	11.025	1.847	11.007	1.866	97.974
S_{10}	11	12.000	1.881	11.977	1.899	98.113

efficiency can be better observed when we deal with large number of genes with extreme intensities which lead to the disturbance of normality assumption of signals.

On the other side, in order to evaluate the performance of both RGX and FGX in a real dataset, we use a one-channel microarray data of a boron toxicity analysis [19], where two different conditions for boron toxicity of barley leaves are compared with a control group. In this analysis, the results are compared with the RMA estimates in terms of the detection of significant genes, fold change at least two, and ROC curve under small and large number of genes. The findings indicate that both FGX and RGX outperform RMA in terms of the control of significant genes and ROC analysis, but they are not more efficient than RMA in the detection of at least 2 fold-changed genes which is one of the strong side of RMA index [20]. Because, in particular, RGX is mostly concentrated on the tails of the density, whereas RMA detects the fold changes around the center of the density under a deterministic approach. Accordingly, when the fold-change is observed under low ratio as found in the boron toxicity analysis, RMA can detect the associated genes better than RGX. Moreover, from the comparative analysis of PAMSAM (partitioning around medoids by using average silhouette width) clustering [21] of fold-changed genes, we observe that the estimates of both FGX and RGX are similar to the RMA's outputs and can produce biologically validated findings [22].

4. Conclusion and Discussion

We have developed an extension of the FGX (frequentist gene expression index) method under the long-tailed symmetric distribution on the logarithmic scale. In inference, we have implemented the modified maximum likelihood method which enables us to solve the intractable likelihood equations and derive the covariances and variances of all estimates. From the analysis of bench-mark data it is seen that the novel estimators are better in the signal detect R^2 and the average R^2, give comparable slopes under different regressions of intensities versus concentration, and still gain from the computational cost while maintaining high sensitivity. Moreover, from the analysis of the simulated data, it is observed that

the strongness of RGX (robust gene expression index) over FGX is clearly seen when the intensities are far from the normality or extreme observations. Therefore, we think that RGX can successively deal with such a high dimensional decision-making problem in inference of the signals, and it enables us to effectively implement the microarray analyses in biochemical studies. The improvement in the estimated signals via RGX can also help us to better reveal the uncertainty in the data by different classification [1] and data mining techniques [3] that we may need during different stages of the biomolecular analyses.

On the other hand, we can improve the performance of RGX in different ways in the sense that the model can be extended by defining signal values with both gene and probe specific, rather than only gene specific values. Additionally, although we assume a constant variance for all probes and genes which seem plausible for Affymetrix spike-in data, it can be constructed under the assumption of gene specific variances [22]. Finally, it is known that the difference between perfect matches PM and mismatches MM values is originated from the base change in the 13th entry of the base sequence. This difference can be also inherently dependent on the annealing temperature between these two sorts of probes and actual annealing temperature of the experiment. Hereby, if this temperature is not equal for all probes on the array, the probe pair can be affected by this difference. This challenge has been discussed in the study of [23], saying that the base pair used on the 13th letter significantly affects the intensities of oligonucleotide. In the study of [24], it is also found that the intensity of PM increases significantly when the PM middle base is a C (Cytosine) or a T (Thymine), whereas the intensity of MM raises considerably when the MM middle base is G (Guanine) or A (Adenine). Considering this distinction coming from the sequence of the base, the PDNN (positional-dependent nearest neighbor) model [25] decomposes the signal in several components according to the formation of RNA-DNA duplexes with many genes. So, similar to that model, we can assign a different weight factor at each base (nucleotide) position on a probe so that different parts of the probe may contribute differently to the stability of the binding.

Acknowledgments

The author would like to thank very much Professor Ernst Wit, Professor Moti L. Tiku, Professor Thomas A. Louis, Professor Carl James Schwarz, the referees, and Professor Gerhard Wilhem Weber for their valuable suggestions which contributed to the improvement of the paper.

References

[1] Z. Volkovich, Z. Barzily, G. W. Weber, D. T. Kiati, R. A. Avros, and R. A. Avros, "An application of the minimal spanning tree approach to the cluster stability problem," *Central European Journal of Operations Research*. In press.

[2] S. Özöğür-Akyüz and G.-W. Weber, "Infinite kernel learning via infinite and semi-infinite programming," *Optimization Methods & Software*, vol. 25, no. 4–6, pp. 937–970, 2010.

[3] G.-W. Weber, Ö. Defterli, S. Z. Alparslan Gök, and E. Kropat, "Modeling, inference and optimization of regulatory networks based on time series data," *European Journal of Operational Research*, vol. 211, no. 1, pp. 1–14, 2011.

[4] M. U. Akhmet, D. Aruğaslan, and E. Yılmaz, "Stability in cellular neural networks with a piecewise constant argument," *Journal of Computational and Applied Mathematics*, vol. 233, no. 9, pp. 2365–2373, 2010.

[5] E. Wit and J. McClure, *Statistics for Microarrays*, John Wiley & Sons, Chichester, UK, 2004.

[6] A.-M. K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green, "BGX: a fully bayesian gene expression index for Affymetrix GeneChip data," *Biostatistics*, vol. 6, no. 3, pp. 349–373, 2005.

[7] X. Liu, M. Milo, N. D. Lawrence, and M. Rattray, "A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips," *Bioinformatics*, vol. 21, no. 18, pp. 3637–3644, 2005.

[8] E. Hubbell, W. M. Liu, and R. Mei, "Robust estimators for expression analysis," *Bioinformatics*, vol. 18, no. 12, pp. 1585–1592, 2002.

[9] Affymetrix, *Statistical Algorithms Description Document*, Affymetrix, Santa Clara, Calif, USA, 2002.

[10] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer, "A model-based background adjustment for oligonucleotide expression arrays," *Journal of the American Statistical Association*, vol. 99, no. 468, pp. 909–917, 2004.

[11] V. Purutçuoğlu and E. Wit, "FGX: a frequentist gene expression index for Affymetrix arrays," *Biostatistics*, vol. 8, no. 2, pp. 433–437, 2007.

[12] M. L. Tiku and A. Akkaya, *Robust Estimation and Hypothesis Testing*, New Age International Ltd., New Delhi, India, 2004.

[13] M. Milo, A. Fazeli, M. Niranjan, and N. D. Lawrence, "A probabilistic model for the extraction of expression levels from oligonucleotide arrays," *Biochemical Society Transactions*, vol. 31, no. 6, pp. 1510–1512, 2003.

[14] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 1, pp. 31–36, 2001.

[15] M. L. Tiku, W. Y. Tan, and N. Balakrishnan, *Robust Inference*, vol. 71, Marcel Dekker, New York, NY, USA, 1986.

[16] G. K. Bhattacharyya, "The asymptotics of maximum likelihood and related estimators based on type II censored data," *Journal of the American Statistical Association*, vol. 80, no. 390, pp. 398–404, 1985.

[17] K. J. Antonellis, Y. D. B. Barclay, M. Elashoff et al., "Optimization of an external standard for the normalization of Affymetrix GeneChip arrays," Tech. Rep., Gene Logic Inc., 2002.

[18] L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed, "A benchmark for Affymetrix GeneChip expression measures," *Bioinformatics*, vol. 20, no. 3, pp. 323–331, 2004.

[19] M. T. Öz, R. Yilmaz, F. Eyidoğan, L. de Graaff, M. Yücel, and H. A. Öktem, "Microarray analysis of late response to boron toxicity in barley (Hordeum vulgare L.) leaves," *Turkish Journal of Agriculture and Forestry*, vol. 33, no. 2, pp. 191–202, 2009.

[20] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[21] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data*, John Wiley & Sons, New York, NY, USA, 1990.

[22] V. Purutçuoğlu, E. Kayış, and G. W. Weber, "Background normalization in Affymetrix arrays and a case study," in *Studies in Computational Intelligence*, Springer, 2011.

[23] F. Naef and M. O. Magnasco, "Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays," *Physical Review E*, vol. 68, no. 1, Article ID 011906, pp. 1–4, 2003.

[24] D. Hekstra, A. R. Taussig, M. Magnasco, and F. Naef, "Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays," *Nucleic Acids Research*, vol. 31, no. 7, pp. 1962–1968, 2003.

[25] L. Zhang, M. F. Miles, and K. D. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnology*, vol. 21, no. 7, pp. 818–941, 2003.