Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 31

5-1-2016

Application of Esscher Transformed Laplace Distribution in Microarray Gene Expression Data

Shanmugasundaram Devika
Christian Medical College, devika@cmcevllore.ac.in

Sebastian George
St. Thomas College, sthottom@gmail.com

Lakshmanan Jeyaseelan
Christian Medical College, ljey@cmcvellore.ac.in

Application of Esscher Transformed Laplace Distribution in Microarray Gene Expression Data

Cover Page Footnote
Shanmugasundaram Devika, M.Sc is a Research scholar in Department of Biostatistics, Christian Medical College, Vellore, India. Email her at: devika@cmcvellore.ac.in. Sebastian George, Ph.D is Professor in Department of Statistics, St. Thomas College, Palai, Kerala, India. Email him at: sthottom@gmail.com. L. Jeyaseelan, Ph.D, FRSS is Professor in Christian Medical College, Vellore, India. Email him at: ljey@cmcvellore.ac.in.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 616-634.

ISSN 1538 − 9472

Application of Esscher Transformed Laplace Distribution in Microarray Gene Expression Data

Shanmugasundaram Devika
Christian Medical College
Vellore, India

Sebastian George
St. Thomas College
Palai, Kerala, India

Lakshmanan Jeyaseelan
Christian Medical College
Vellore, India

Microarrays allow the study of the expression profile of hundreds to thousands of genes simultaneously. These expressions could be from treated samples and the healthy controls. The Esscher transformed Laplace distribution is used to fit microarray expression data as compared to Normal and Laplace distributions. The Maximum Likelihood Estimation procedure is used to estimate the parameters of the distribution. R codes are developed to implement the estimation procedure. A simulation study is carried out to test the performance of the algorithm. AIC and BIC criterion are used to compare the distributions. It is shown that the fit of the Esscher transformed Laplace distribution is better as compared to Normal and standard Laplace distributions.

Keywords: Esscher transformed Laplace distribution, Normal distribution, Laplace distribution, Microarray gene expression, Maximum Likelihood estimation

Introduction

Microarrays allow the researcher to investigate the expressions of thousands of genes simultaneously under various condition of the biological process. These conditions could be samples from cancer tumor and healthy controls. This method measures the intensity of the fluorescence after hybridization and then expression profiles are compared between two different samples of Complementary DNA (cDNA) colored with different dyes, Red (for diseased) and Green (for healthy control). Hence this method allows us to study the relative gene expression in two different samples. The statistical methods that have been developed to analyze the gene expression data over the decades depend heavily on the distribution of the gene expression data.

Ms. Devika is a PhD Research Scholar. Email her at: devika@cmcvellore.ac.in. Dr. George is a Professor in the Department of Statistics. Email him at: sthottom@gmail.com. Dr. Jeyaseelan is a Professor. Email him at: ljey@cmcvellore.ac.in.

The gene expression data, after normalization, usually has a heavier tail as compared to normal distribution. That is, most of the mass at the center with a sharp peak with varying asymmetry. Researchers have used several densities to model gene expression data. Densities of Poisson, exponential, and logarithmic series were used (Kuznetsov, 2001). An error distribution of gene expression datasets was approximated by two distributions by taking log-normal in the bulk of microarray spot intensities and a power law in the tails (Hoyle, Rattray, Jupp, & Brass, 2002). The gene expression was also fitted by using an asymmetric Laplace distribution (Purdom & Holmes, 2005). However, in order to take outliers into account, the Cauchy distribution has been used for estimating gene expressions using data from multiple-laser scans (Khondoker, Glasbey, & Worton, 2006), and the Laplace mixture model was introduced as a long tailed alternative to the normal distribution (Bhowmick, Davison, Goldstein, & Ruffieux, 2006).

Recently, asymmetric type II compound Laplace density (Punathumparambath, Kulathinal, & George, 2012) was introduced for the analysis of gene expression data which was asymmetric version of type II compound Laplace distribution and a generalization of asymmetric Laplace distribution. The four parameter probability distribution provided an additional degree of freedom to capture the characteristic feature of the microarray data. Based on the above review, the microarray data with thousands of genes show asymmetry and most of the mass at the middle as large proportion of genes are not differently expressed. Therefore the log ratio of the intensities have a tendency to cluster around a single point and with the presence of outliers. Hence it may not be appropriate to summarize such pattern with mean, variance, etc.

In the current study, new class of asymmetric Laplace distribution is proposed for the analysis of log ratios of measured gene expression data across genes through Esscher transformation, namely Esscher transformed Laplace (ETL) distribution proposed in George and George (2012). It is a sub-class of one parameter exponential family and an alternative to various types of asymmetric Laplace distributions given in Kotz, Kozubowski, and Podgórski (2001). If all the genes on one array are considered as separate independent observations, the distribution of the log-ratio of the expression values is well approximated by the asymmetric nature of the ETL distribution. Moreover modeling distribution with single parameter would be a feasible approach as compared to distribution such as asymmetric type II compound Laplace distribution with four parameters. This paper presents the analysis of microarray gene expression data using the ETL distribution. The paper is organized as follows: First we describe the overview of ETL distribution, followed by a simulation study. Next Normal, Laplace, and ETL

distributions were fitted to gene expression data and compared. Finally the paper ends with conclusion.

Methods

Overview of Esscher Transformed Laplace Distribution

The ETL distribution was proposed in George and George (2012) and George (2011). A random variable X is said to follow Esscher transformed Laplace distribution with parameter (θ) if its probability distribution function (pdf) is given by

$$\mathrm{f}\left(x\,|\,\theta\right)=\frac{1-\theta^2}{2}\begin{cases}\exp\left[x\left(1+\theta\right)\right], & x<0\\ \exp\left[-x\left(1-\theta\right)\right], & x\geq 0\end{cases} \tag{1}$$

where θ is called the Esscher parameter and $\theta \in$ (-1, 1). This pdf can also be expressed as

$$\mathrm{f}\left(x\,|\,\theta\right)=\frac{1-\theta^2}{2}\exp\left[\theta x-|x|\right];\ -\infty<x<\infty,\ \theta\in\left(-1,1\right) \tag{2}$$

Thus the ETL distribution is a regular one parameter exponential family and a subclass of the family of asymmetric Laplace (AL) distributions proposed in Kotz et al. (2001). These kinds of distribution are more appropriate for modeling financial datasets as this allows for asymmetry, peakedness and tailed heaviness than normal distribution (George & George, 2013).

The cumulative distribution function (cdf) of the ETL distribution is given by

$$\mathrm{F}\left(x\,|\,\theta\right)=\begin{cases}\dfrac{\left(1-\theta\right)}{2}\exp\left[x\left(1+\theta\right)\right], & x<0\\ \dfrac{1-\theta}{2}+\dfrac{1+\theta}{2}\left(1-\exp\left[-x\left(1-\theta\right)\right]\right), & x\geq 0\end{cases} \tag{3}$$

for $\theta \in$ (-1, 1). When θ = 0, we get the classical Standard Laplace (0, 1) distribution.

Figure 1 represents the densities of the ETL distribution. When $\theta \in$ (-1, 0) the distribution is left skewed and $\theta \in$ (0, 1) the distribution is right skewed. From

Figure 1 we can see that the ETL distribution has heavier tails than the normal distribution, meaning that there is more probability of extreme values than under a normal distribution. In addition, the ETL distribution concentrates more probability in the center than a normal distribution. It is also clear from Figure 1 that the shape of the ETL distribution is nearly similar to the AL distribution but the later does not belong to one parametric exponential family whereas the former does.

The characteristic function of the AL (μ, σ) with parameters $\mu \in \mathbb{R}$ and $\sigma \geq 0$ and ETL (θ) distributions are given by

$$\varphi_{\mathrm{X}}(t) = \left(1 + \sigma^2 t^2 - i\mu t\right)^{-1}$$

and

$$\varphi_{\mathrm{X}}(t) = \left(1 + \frac{t^2}{1-\theta^2} - \frac{2it\theta}{1-\theta^2}\right)^{-1}$$

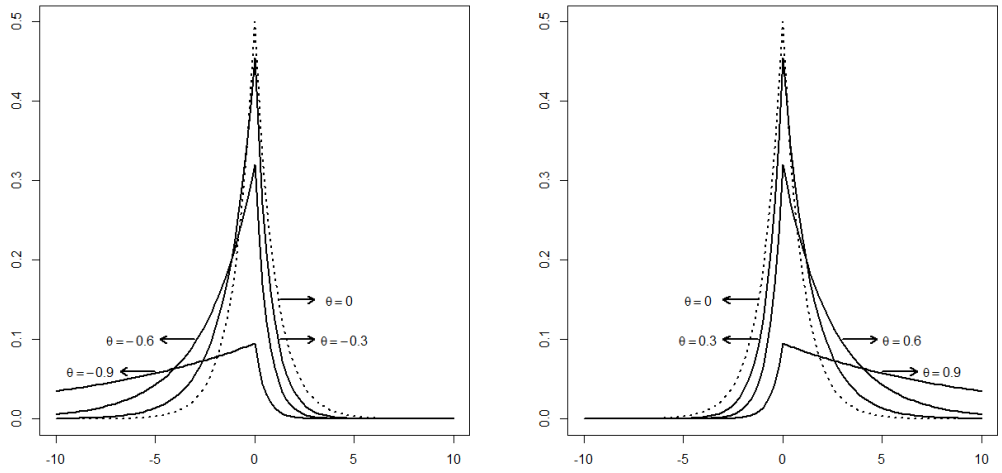

Figure 1. Densities of the Esscher transformed Laplace distribution for various choices of parameter θ.

Hence ETL (θ) is a special case of the AL (μ, σ) distribution with $\mu = 2\theta/(1-\theta^2)$ and $\sigma^2 = 1/(1-\theta^2)$. The mean E($x$) and variance Var($x$) of the ETL distribution are given by

$$\mathrm{E}(x) = \frac{2\theta}{1-\theta^2}; \quad \mathrm{Var}(x) = \frac{2\left(1+\theta^2\right)}{\left(1-\theta^2\right)}$$

The α^{th} quantile of the ETL (θ) distribution for simulation purpose in the later section is given by

$$q_a = \begin{cases} \dfrac{1}{1+\theta}\log\left[\dfrac{2}{1-\theta}\alpha\right], & \alpha \in \left(0, \dfrac{1-\theta}{2}\right) \\ \dfrac{1}{1+\theta}\log\left[\dfrac{2}{1+\theta}(1-\alpha)\right], & \alpha \in \left(\dfrac{1-\theta}{2}, 1\right) \end{cases} \tag{4}$$

The parameter of the ETL distribution can be obtained either by the method of maximum likelihood (MLE) or by the method of moments. Let $x_1, x_2, \ldots, x_n$ be an independent identically distributed (i.i.d) random variable from the ETL (θ) distribution with density from equations (1) or (2). The likelihood function is then written as

$$\log \mathrm{L}(X:\theta) = \mathrm{n}\log\left(\frac{1-\theta^2}{2}\right) + \theta\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}|x_i|$$

and the first derivative with respect to the parameter θ is

$$\frac{\partial \log \mathrm{L}}{\partial\theta} = \frac{-2n\theta}{1-\theta^2} - \sum_{i=1}^{n} x_i$$

The MLE of parameter θ is obtained by solving the score function

$$\frac{\partial \log \mathrm{L}}{\partial\theta} = 0$$

so that

$$\hat{\theta} = \frac{-1 \pm \sqrt{1+\bar{x}^2}}{\bar{x}}$$

provided that $\theta \in (-1, 1)$.

By introducing the location parameter (μ) and scale parameter (σ) in the ETL distribution, the pdf and cdf of the ETL (θ, μ, σ) distribution is given as follows:

$$\mathrm{f}\left(x \mid \theta, \mu, \sigma\right) = \begin{cases} \dfrac{\left(1-\theta^2\right)}{2\sigma} \exp\left[\left(\dfrac{x-\mu}{\sigma}\right)\left(1+\theta\right)\right], & x < \mu \\ \dfrac{\left(1-\theta^2\right)}{2\sigma} \exp\left[\left(\dfrac{\mu-x}{\sigma}\right)\left(1-\theta\right)\right], & x \geq \mu \end{cases} \tag{5}$$

and

$$\mathrm{F}\left(x \mid \theta, \mu, \sigma\right) = \begin{cases} \dfrac{\left(1-\theta\right)}{2} \exp\left[\left(\dfrac{x-\mu}{\sigma}\right)\left(1+\theta\right)\right], & x < \mu \\ 1 - \dfrac{\left(1+\theta\right)}{2} \exp\left[\left(\dfrac{\mu-x}{\sigma}\right)\left(1-\theta\right)\right], & x \geq \mu \end{cases} \tag{6}$$

where $\theta \in (-1, 1)$, $\mu \in \mathbb{R}$, and $\sigma > 0$.
The mean E(x) and variance Var(x) of the ETL with location μ and scale parameter σ are given by

$$\mathrm{E}\left(x\right) = \mu + \frac{2\theta\sigma}{1-\theta^2}$$

$$\mathrm{Var}\left(x\right) = \frac{2\sigma^2\left(1+\theta^2\right)}{\left(1-\theta^2\right)^2}$$

The α^{th} quantile of the ETL (θ, μ, σ) distribution is

$$q_a = \begin{cases} \mu + \dfrac{\sigma}{1+\theta}\log\left(\dfrac{2}{1-\theta}q\right), & q \in \left(0, \dfrac{1-\theta}{2}\right) \\ \mu - \dfrac{\sigma}{1-\theta}\log\left(\dfrac{2}{1+\theta}(1-q)\right), & q \in \left(\dfrac{1-\theta}{2}, 1\right) \end{cases} \tag{7}$$

The parameters θ, μ, and σ of the ETL distribution were obtained by maximization of the likelihood function in R software (R Development Core Team, 2014) using optim function with BFGS (Broyden, Fletcher, Goldfarb, and Shanno) algorithm. The standard error (SE) of the respective parameters were obtained by inverting the Fisher information matrix at the maximum likelihood estimates. As this was a methodological study which used open source data, IRB clearance was not necessary.

Data Simulation

A simulation experiment is executed to study the functioning of the estimation algorithm for various arbitrary values of the parameters of the ETL (θ, μ, σ) distribution. We created 1000 datasets each with sample of size n = 2000 from the ETL distribution by fixing the Esscher parameter θ = (-0.5, 0, 0.5), location parameter μ = (-0.5, -0.2, 0.3, 0.9), and scale parameter σ = (0.5, 0.75, 1, 1.5) by using an inverse transform sampling procedure. Then the maximum likelihood estimates of the parameters are obtained as mentioned above by using R statistical software. Table 1 represents the results of the simulation study performed by using 1000 different data sets. It is apparent that the estimation procedure works well for different choices of parameters and the sample standard deviation are in accordance with the asymptotic standard error obtained using maximum likelihood estimate. However the difference increases with increase in the σ values. We also checked the convergence of the estimation procedure for various choices of parameter values with different initials and the algorithm works satisfactorily well for several alternatives.

Results

Analysis of Microarray gene expression data

The ETL distribution was applied to three different microarray datasets (Swirl, E. coli, and Tumor) from published microarray experiments. The first data set Swirl

zebrafish experiment is included as part of the marray package in R software (Dudoit & Yang, 2002). This data is provided by Katrin Wuennenberg-Stapleton from the Ngai Lab at UC Berkeley (2001). Swirl is a point mutant in the vertebrates. In order to access the mutational status, zebrafish was taken as a model organism. The aim of the experiment was to find genes which were differentially expressed between mutant and wild type zebrafish. The cDNA from wild type mutant was labelled using Cy3 dyes and the swirl mutant with Cy5. There were totally four replicates (Swirl.1,...., Swirl.4) and the target cDNA was hybridized to microarrays containing 8,448 probes, including 768 control spots. The raw dataset was first log transformed to base 2 and normalized using a print tip group Lowess smoothing technique (locally weighted linear regression method) (Cleveland & Devlin, 1988) and with quantile normalization procedure. This method is widely used in microarray experiments as this removes the intensity dependence in $\log_2(R_i/G_i)$ values, where R_i is the red dye intensity (Cy3) and G_i (Cy5) is the green dye intensity for the i^{th} gene (Yang et al., 2002). The same dataset was used to fit asymmetry Laplace distribution in Purdom and Holmes (2005).

The next dataset, E. coli, was a two channel microarray experiment conducted to compare gene expression profiles of wild strain with mutant strain and was provided by Bernstein, Lin, Cohen, and Lin-Chao (2004). The dataset contained information on 5128 genes with six arrays. mRNA extracted from wild strain was labeled with Cy5 (Green) and the mutant strain with Cy3 (Red). The E. coli data was also normalized using Lowess technique and the quantile normalization procedure and then the log differences was taken as gene expression measurement. The third dataset Tumor microarray experiment was carried on to compare the functioning of gene expression of ovarian tumor cells as compared to normal cells. This study involved six samples from normal cells and six from ovarian tumor cells on 34,742 genes. We transformed the data using log function with base 2 and then we used Lowess and quantile normalization procedure as earlier.

Gaussian, Laplace, and ETL distributions were fitted to log transformed normalized gene expression measurements $\log_2(R_i/G_i)$ for the three datasets. The parameters of the Gaussian (μ, σ^2), Laplace (μ, σ), and ETL (θ, μ, σ) distributions were estimated using maximum likelihood estimation method and their corresponding standard errors. In Table 2, results for two arrays from each dataset are presented, and the rest are given in the supplementary Table 4.

Table 1. Simulation study – maximum likelihood estimates of θ, μ, and σ for various choices of parameters

θ	σ	μ	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\mu}$	SE($\hat{\theta}$)	SD($\hat{\theta}$)	SE($\hat{\sigma}$)	SD($\hat{\sigma}$)	SE($\hat{\mu}$)	SD($\hat{\mu}$)
-0.5	0.50	-0.5	-0.4999	0.5001	-0.5001	0.0177	0.0199	0.0163	0.0171	0.0147	0.0190
	0.75	-0.2	-0.4999	0.7500	-0.2000	0.0173	0.0199	0.0241	0.0257	0.0204	0.0286
	1.00	0.3	-0.4999	1.0001	0.3000	0.0170	0.0199	0.0319	0.0342	0.0259	0.0382
	1.50	0.9	-0.5000	1.5000	0.9000	0.0166	0.0199	0.0472	0.0514	0.0358	0.0572
0.0	0.50	-0.5	0.0003	0.5000	-0.5002	0.0210	0.0227	0.0112	0.0110	0.0135	0.0168
	0.75	-0.2	0.0003	0.7500	-0.2004	0.0203	0.0227	0.0168	0.0165	0.0187	0.0252
	1.00	0.3	0.0003	1.0000	0.2995	0.0199	0.0228	0.0224	0.0220	0.0235	0.0337
	1.50	0.9	0.0003	1.5001	0.8992	0.0197	0.0228	0.0336	0.0329	0.0337	0.0505
0.5	0.50	-0.5	0.5015	0.4987	-0.5017	0.0180	0.0199	0.0165	0.0174	0.0150	0.0199
	0.75	-0.2	0.5015	0.7481	-0.2025	0.0172	0.0200	0.0241	0.0261	0.0203	0.0300
	1.00	0.3	0.5015	0.9974	0.2966	0.0171	0.0199	0.0320	0.0347	0.0261	0.0399
	1.50	0.9	0.5016	1.4961	0.8949	0.0166	0.0200	0.0472	0.0521	0.0362	0.0599

Figures 2-3 represent the box plots of intensities of Swirl, E. coli, and Tumor datasets before and after normalization. It is clear from Figures 2-3 that, after normalization, each distribution of the gene expression has a similar shape and exhibits heavier tails with a certain degree of asymmetry as compared to a Gaussian distribution. The left side of Figures 4-9 and supplementary Figures 10-19 shows the histogram super imposed with ETL (θ, μ, σ), Laplace (μ, σ) and Gaussian (μ, σ^2) distributions, where the parameters of these distributions were obtained by the maximum likelihood estimation procedure. By comparing these densities, ETL (θ, μ, σ) captures the asymmetric nature of the data with peaked concentration in the middle and heavy tail.

It can be seen from Table 2 that the Esscher parameter (θ) for arrays Swirl.1 and Swirl.3 are greater than 0 (right skewed) and for all the other arrays the parameter (θ) is smaller than 0 (left skewed). Though the level of skewness in all the arrays of the datasets is not very large, they are different from 0. It is also noted that the maximum likelihood estimate of parameter σ of the ETL and Laplace distributions are approximately equal.

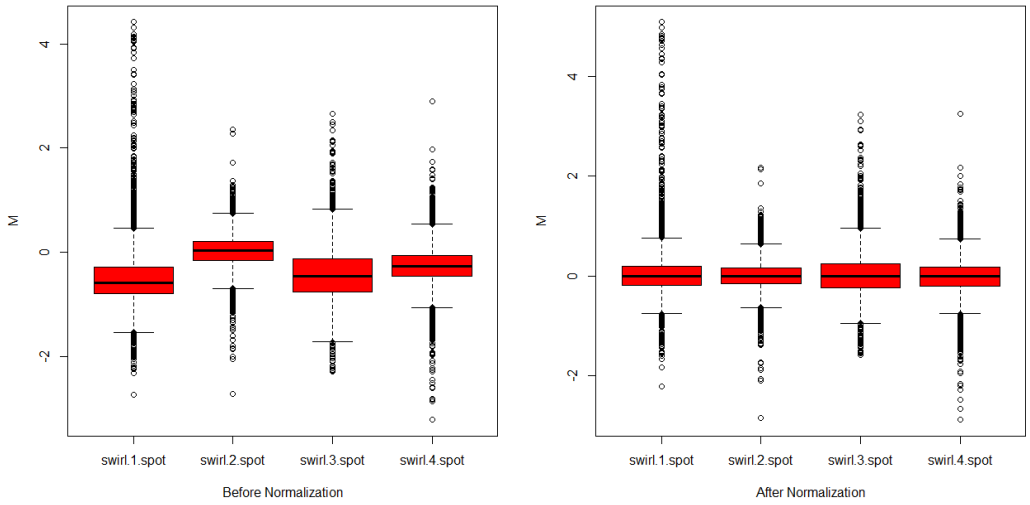

Figure 2. Boxplot of intensities from Swirl zebrafish microarray experiment, before and after normalization.

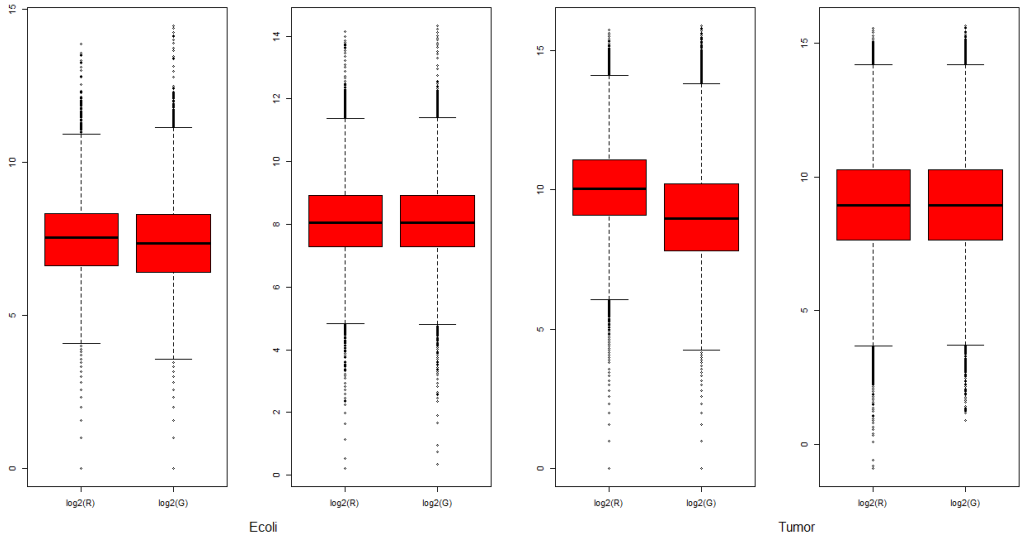

Figure 3. Boxplot of intensities of Red and Green arrays of Ecoli and Tumor microarray experiments, before and after normalization.

Table 2. Microarray data analysis – maximum likelihood estimates and the asymptotic standard error for Esscher transformed Laplace, Laplace, and Normal distributions.

	Swirl.1	Swirl.3	Ecoli.1	Ecoli.2	Tumor.3	Tumor.5
Esscher						
θ	0.24(0.0128)	0.23(0.0111)	-0.090(0.0188)	-0.160(0.0159)	-0.080(0.0063)	-0.080(0.0064)
σ	0.26(0.0034)	0.30(0.0036)	0.330(0.0047)	0.430(0.0065)	0.710(0.0039)	0.660(0.0036)
μ	-0.09(0.0058)	-0.10(0.0052)	0.060(0.0106)	0.140(0.0112)	0.110(0.0072)	0.110(0.0069)
Laplace						
μ	-0.01(0.0035)	-0.01(0.0038)	0.020(0.0060)	0.050(0.0082)	0.040(0.0047)	0.040(0.0043)
σ	0.29(0.0031)	0.32(0.0035)	0.330(0.0046)	0.450(0.0063)	0.710(0.0038)	0.660(0.0036)
Gaussian						
μ	0.05(0.0052)	0.04(0.0047)	0.002(0.0068)	0.002(0.0092)	0.005(0.0054)	0.005(0.0051)
σ	0.23(0.0035)	0.19(0.0029)	0.240(0.0047)	0.430(0.0085)	1.030(0.0078)	0.890(0.0068)

One of the graphical procedures to compare the probability distribution Quantile-Quantile plot (Q-Q plot) is shown in the right side of Figures 4-9 and supplementary Figures 10-19. This is obtained by plotting the theoretical quantiles against sample quantiles. This plot is more useful as this better emphasizes the fit of the distributions in the tail region. It is indicated in Figures 4-9 that the ETL (θ, μ, σ) distribution fits to the data well as compared to other two distributions, especially when (θ) is significantly greater than 0 (right skewed) for Swirl.1 and 3 and smaller than 0 (left skewed) for all the other arrays. The supplementary Figures 10-19 indicate that, when $\theta \approx 0$, the performance of both the Laplace and ETL distributions are almost similar but still better than Gaussian distribution. Other than with few outliers, the fit of the ETL distribution is greatly improved as compared to the other distributions considered, though all the three seem to describe the middle region of the data rather similarly.

A numerical evaluation of model comparison was done by using Akaike's Information Criterion (AIC) (Akaike, 1998) and Bayesian Information Criterion (BIC) (Schwarz, 1978) as the later take into account of the sample size. The formula for AIC and BIC are given by

$$\mathrm{AIC} = -2\log\left(\mathrm{L}_g\left(\hat{\theta} \mid x_1, \ldots, x_n\right)\right) + 2K$$

and

$$\mathrm{BIC} = -2\log\left(\mathrm{L}_g\left(\hat{\theta} \mid x_1, \ldots, \mathrm{x}_\mathrm{n}\right)\right) + K\log(n)$$

where K is the number of parameters being estimated, L is the likelihood function of the model g, $\hat{\theta}$ is the maximum likelihood estimate of the parameters of model g, and n is the sample size. Given the different models, the one with smaller AIC/BIC fits the data better than the one with the larger AIC/BIC, where the conclusion from AIC and BIC goes hand in hand in most of the cases. AIC and BIC values of the three distributions, ETL (θ, μ, σ), Laplace (μ, σ), and Gaussian (μ, σ^2) are given in Table 3 and supplementary Table 5. The ETL (θ, μ, σ) distribution had a lower AIC/BIC values for all the sample arrays shown in Table 3. Hence the ETL distribution shows an improvement in the model fit as compared to other distributions. However, when there is an absence of asymmetry ($\theta \approx 0$) the values of AIC/BIC for the ETL distribution are nearly equal to the Laplace distribution. This feature has been seen in the arrays of Swirl.2, Ecoli.4, Ecoli.5, Ecoli.6 and Tumor.2 in supplementary Table 5, which shows a similar performance of ETL and Laplace distributions.

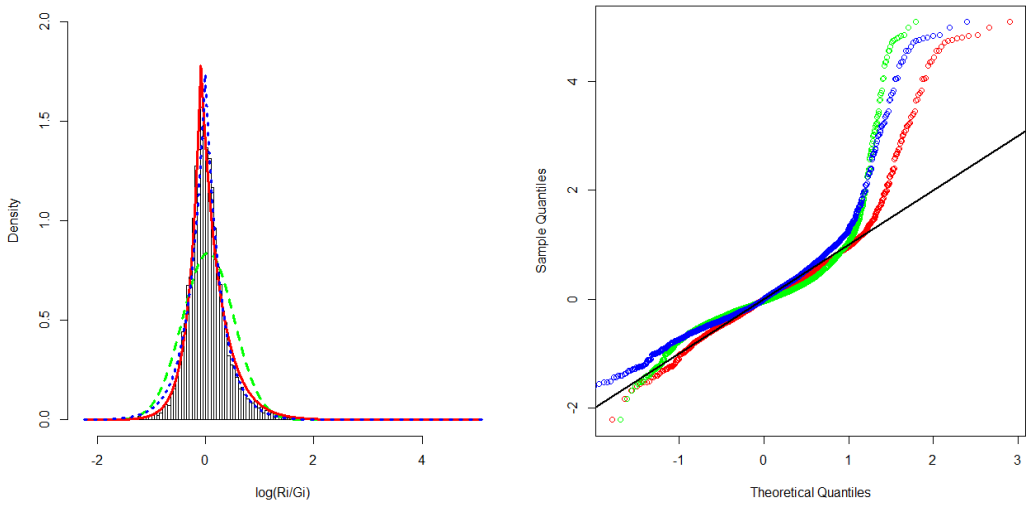

Figure 4. Left: Histogram of Swirl.1 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

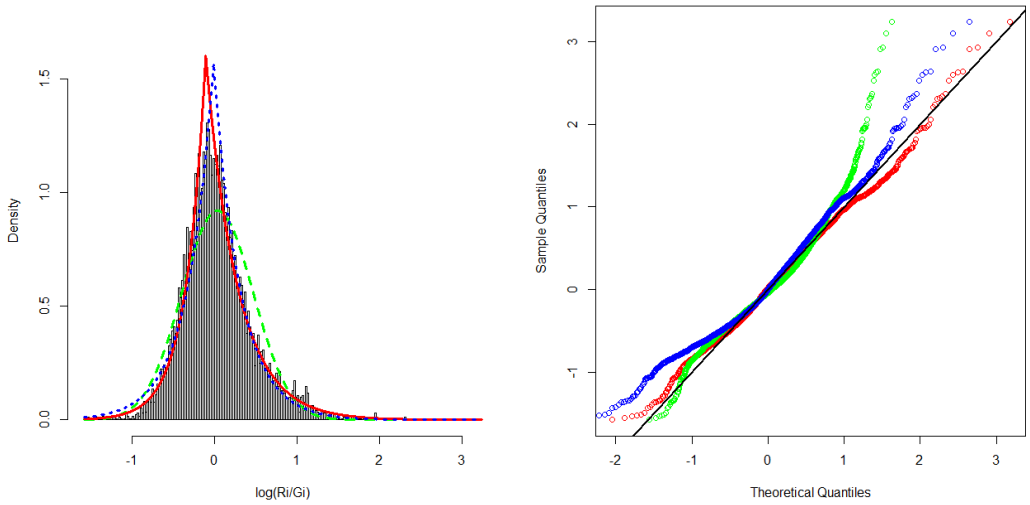

Figure 5. Left: Histogram of Swirl.3 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

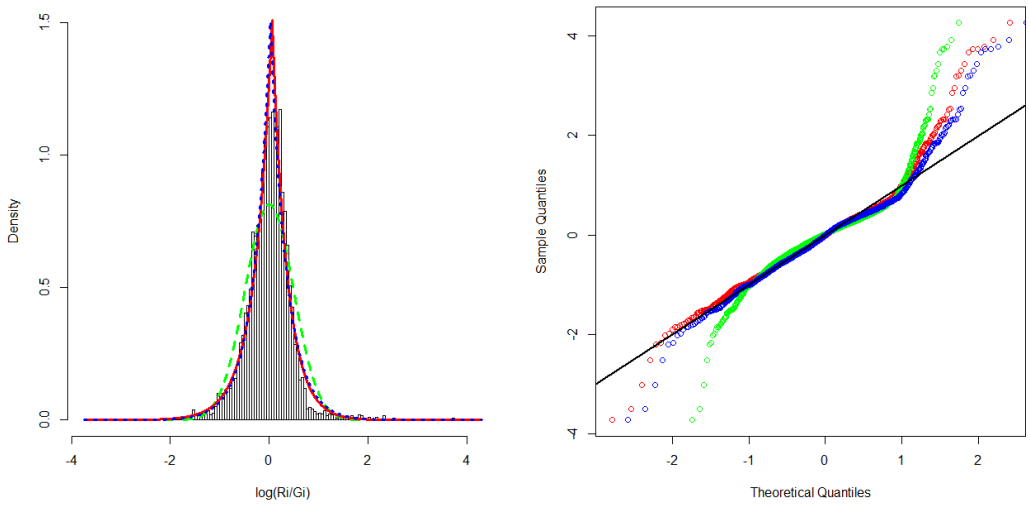

Figure 6. Left: Histogram of Ecoli.1 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

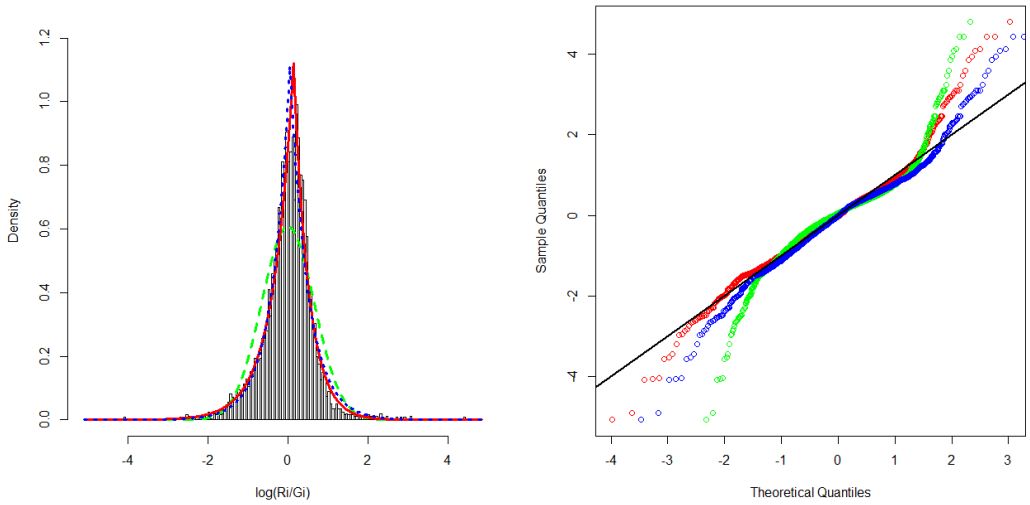

Figure 7. Left: Histogram of Ecoli.2 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

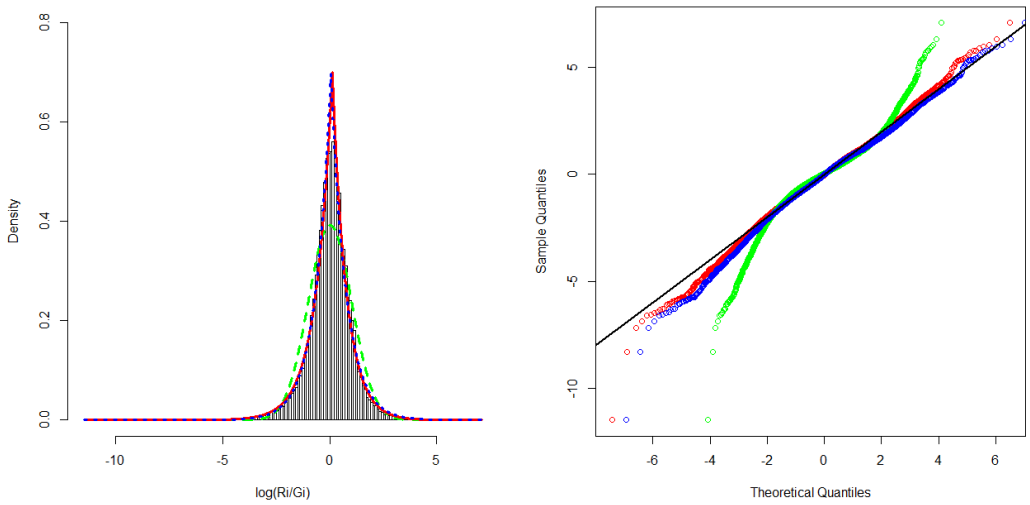

Figure 8. Left: Histogram of Tumor.3 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

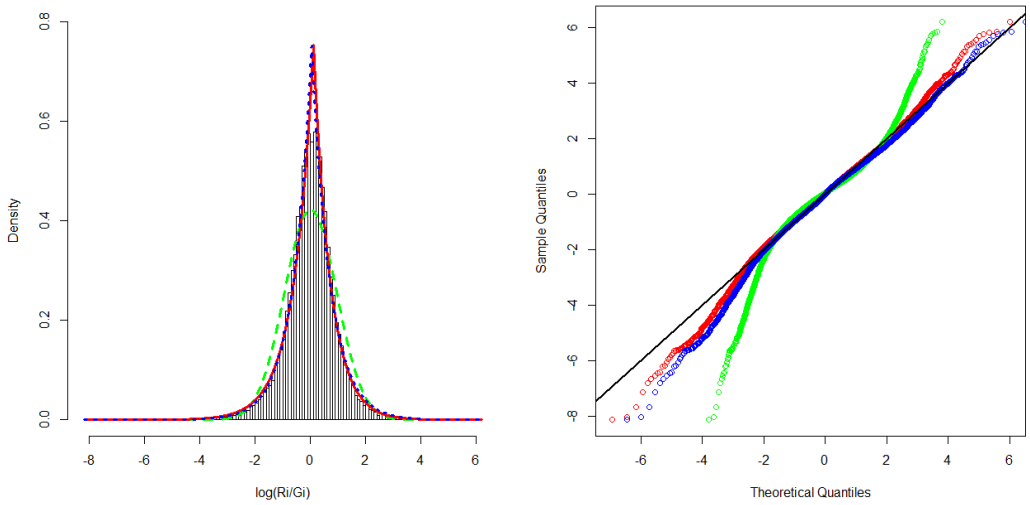

Figure 9. Left: Histogram of Tumor.5 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

Table 3. Comparison of AIC and BIC of Esscher transformed Laplace, Laplace, and Normal distributions.

	Swirl.1		Swirl.3		Ecoli.1		Ecoli.2		Tumor.3		Tumor.5	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Esscher	7125	7146	8942	8963	6023	6043	9084	9104	94157	94183	89044	89069
Laplace	7549	7563	9245	9259	6045	6058	9167	9180	94301	94318	89200	89217
Gaussian	11406	11420	9855	9869	7234	7247	10248	10261	99634	99651	94568	94585

Conclusion

In the two channel microarray experiments, for which the ETL distribution was fitted, gave a reasonable fit to the gene expression data and greatly improved upon the normal distribution and as an alternative to Laplace distribution. The ETL (θ, μ, σ) can be a better model for gene expression data as they are asymmetric, heavy tailed, and with bulk mass in the middle of the distribution and which does not follow any of the classical symmetric distributions such as Normal, Laplace etc., Esscher transformed Laplace distribution is simple to use distribution which belongs to regular exponential family captures all the features as mentioned above

of the gene expression measurement. In this distribution, the asymmetry is determined by using Esscher parameter (θ) along with the location (μ) and scale (σ) parameters. This distribution is more flexible and belongs to the special case of AL distribution and is also easily tractable for statistical inference. Simulating observations from the ETL distribution is also possible by inverting the cumulative distribution function.

The microarray gene expression data has been modeled using different densities by several authors. AL distribution was introduced in Purdom and Holmes (2005) in the analysis of gene expression data to capture the peak at the center as well as the asymmetry in the distribution. The Laplace mixture model as a long tailed alternative to the normal distribution in identifying differentially expressed genes in microarray experiments was introduced in Bhowmick et al. (2006). The Cauchy distribution was applied in Khondoker et al. (2006) in modeling microarray experiments which can estimate gene expressions by taking the outliers into account. Asymmetric type II compound Laplace distribution in the analysis of microarray gene expression data was introduced in (Punathumparambath et al., 2012). The same author has proposed a family of skew-slash distributions generated by normal kernel (Punathumparambath, 2011), two compound mixture Gaussian models (Punathumparambath, George, & V. M., 2011), skew-slash distributions generated by the Cauchy kernel (Punathumparambath, 2013), skew-slash t and skew-slash Cauchy distributions (Punathumparambath, 2012b), and asymmetric slash Laplace distribution (Punathumparambath, 2012a) for modeling gene expression data.

The ETL distribution was used in modeling microarray data as an alternative to normal and Laplace distributions. From Figures 4-9 and supplementary Figures. 10-19, we can see that the ETL distribution fits the tail region better as compared to other two distributions. This is also evident in the reduction in AIC/BIC values for the ETL distribution as compared to the normal and Laplace distributions. The ETL belongs to exponential family of distributions and is also a generalization of the AL distribution. The main motive of applying different distributions to microarray gene expression data is to capture the asymmetry and peakedness because a large proportion of genes are not differentially expressed, the log ratio of the intensities have tendency to cluster around a single point, and the presence of outliers (Punathumparambath et al., 2012). This distribution is already been applied in George and George (2013) to financial data modeling and web server data, and it was shown that the model fit was better as compared to other distributions.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (199-213). New York: Springer New York.

Bernstein, J. A., Lin, P.-H., Cohen, S. N., & Lin-Chao, S. (2004). Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America, 101*(9), 2758-2763. doi: 10.1073/pnas.0308747101

Bhowmick, D., Davison, A. C., Goldstein, D. R., & Ruffieux, Y. (2006). A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics, 7*(4), 630-641. doi: 10.1093/biostatistics/kxj032

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*(403), 596-610. doi: 10.2307/2289282

Dudoit, S. & Yang, Y. H. (2002). *marrayClasses package: Classes and methods for cDNA microarray data*. Retrieved from http://www.bioconductor.org/

George, D. (2011). *A class of heavy-tailed distributions and their applications* (Unpublished doctoral thesis). Mahatma Gandhi University, Kottayam, Kerala, India.

George, D., & George, S. (2011). Application of Esscher transformed Laplace distribution in web server data. *Journal of Digital Information Management, 9*(1), 19-26.

George, D., & George, S. (2013). Marshall–Olkin Esscher transformed Laplace distribution and processes. *Brazilian Journal of Probability and Statistics, 27*(2), 162-184. doi: 10.1214/11-BJPS163

George S, & George, D. (2012). Esscher transformed Laplace distribution and its applications. *Journal of Probability and Statistical Science, 10*(2), 135-152.

Hoyle, D. C., Rattray, M., Jupp, R., & Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics, 18*(4), 576-584. doi: 10.1093/bioinformatics/18.4.576

Khondoker, M. R., Glasbey, C. A., & Worton, B. J. (2006). Statistical estimation of gene expression using multiple laser scans of microarrays. *Bioinformatics, 22*(2), 215-219. doi: 10.1093/bioinformatics/bti790

Kotz, S., Kozubowski, T. J., & Podgórski, K. (2001). *The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance*. Boston: Birkhäuser.

Kuznetsov, V. A. (2001). Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. *EURASIP Journal on Advances in Signal Processing, 2001*(4), 285-296. doi: 10.1155/S1110865701000294

Punathumparambath, B. (2011). A new family of skewed slash distributions generated by the normal kernel. *Statistica, 71*(3), 345-353. doi: 10.6092/issn.1973-2201/3618

Punathumparambath, B. (2012a). The multivariate asymmetric slash Laplace distribution and its applications. *Statistica, 72*(2), 235-249. doi: 10.6092/issn.1973-2201/3645

Punathumparambath, B. (2012b). The multivariate skew-slash t and skew-slash Cauchy distributions. *Model Assisted Statistics and Applications, 7*(1), 33-40. doi: 10.3233/MAS-2011-0199

Punathumparambath, B. (2013). A new family of skewed slash distributions generated by the Cauchy kernel. *Communications in Statistics - Theory and Methods, 42*(13), 2351-2361. doi: 10.1080/03610926.2011.599508

Punathumparambath, B., George, S., & V. M., K. (2011). Statistical techniques for microarray technology. *Journal of Informatics and Mathematical Sciences, 3*(3), 257-275.

Punathumparambath, B., Kulathinal, S., & George, S. (2012). Asymmetric type II compound Laplace distribution and its application to microarray gene expression. *Computational Statistics & Data Analysis, 56*(6), 1396-1404. doi: 10.1016/j.csda.2011.10.026

Purdom, E., & Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology, 4*(1), . doi: 10.2202/1544-6115.1070

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464. doi: 10.1214/aos/1176344136

University of California, Berkeley (2001). *Swirl experimental data* [Data set]. Provided by the Ngai Lab.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., & Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method

addressing single and multiple slide systematic variation. *Nucleic Acids Research, 30*(4), e15. doi: 10.1093/nar/30.4.e15