The International Journal of Biostatistics

Volume 5, *Issue* 1 2009 *Article* 16

Modelling and Assessing Differential Gene Expression Using the Alpha Stable Distribution

Diego Salas-Gonzalez* Ercan E. Kuruoglu†

Diego P. Ruiz‡

*University of Granada, dsalas@ugr.es

†ISTI-CNR, ercan.kuruoglu@isti.cnr.it

‡University of Granada, druiz@ugr.es

Modelling and Assessing Differential Gene Expression Using the Alpha Stable Distribution*

Diego Salas-Gonzalez, Ercan E. Kuruoglu, and Diego P. Ruiz

Abstract

After normalization, the distribution of gene expressions for very different organisms have a similar shape, usually exhibit heavier tails than a Gaussian distribution, and have a certain degree of asymmetry. Therefore, this distribution has been modeled in the literature using different parametric families of distributions, such the Asymmetric Laplace or the Cauchy distribution. Moreover, it is known that the tails of spot-intensity distributions are described by a power law and the variance of a given array increases with the number of genes. These features of the distribution of gene expression strongly suggest that the alpha-stable distribution is suitable to model it.

In this work, we model the error distribution for gene expression data using the alpha-stable distribution. This distribution is tested successfully for four different datasets. The Kullback-Leibler, Chi-square and Hellinger tests are performed to compare how alpha-stable, Asymmetric Laplace and Gaussian fit the spot intensity distribution. The alpha-stable is proved to perform much better for every array in every dataset considered.

Furthermore, using an alpha-stable mixture model, a Bayesian log-posterior odds is calculated allowing us to decide whether a gene is differently expressed or not. This statistic is based on the Scale Mixture of Normals and other well known properties of the alpha-stable distribution. The proposed methodology is illustrated using simulated data and the results are compared with the other existing statistical approach.

KEYWORDS: microarray gene expression, alpha-stable distribution, mixture model

---

*This work was partially supported by the project TEC2007-68030-C02-02/TCM of the MCyT of Spain, the PETRI project DENCLASES (PET2006-0253) and TEC2008-02113 of the Spanish MEC and the Excellence Projects (TIC-03269 and TIC-02566) of the Consejería de Innovación Ciencia y Empresa (Junta de Andalucía, Spain). The first author performed part of the work while at ISTI-CNR of Pisa.

1 Introduction

DNA microarray has been established as a powerful tool to study the RNA expression levels of thousands of genes simultaneously under different conditions. Namely, these experiments compare two different samples of cDNA dyed with different colours (red and green) by the means of the fluorescence intensity measured in the microarray after hybridization. This methodology allows us to compare a large amount of information simultaneously in order to identify and quantify the genes which are differentially expressed.

It is well known that the independence assumption between genes is not true, but many works that identify differential expression in microarray are based on this assumption Lonnstedt & Speed (2002); Gottardo *et al.* (2003); Bhowmick *et al.* (2006). Most of the approaches based on Bayesian statistical methods assume independence between genes and Gaussian distribution as a device to obtain an analytic formula. However, the distribution of gene expression, also known as the error distribution for gene expression data, has also been modelled under different approaches.

In Kuznetsov (2001), this distribution is modelled using different classes of skewed probability functions such Poisson, exponential, logarithmic series and Pareto-like distribution. The results are shown only for the Pareto-like distribution and it is claimed that this distribution fits the empirical gene expression distribution better than do the other distributions. In Hoyle *et al.* (2002), a wide range of datasets are analyzed empirically and the error distribution is approximated by two distributions: a log-normal in the bulk of microarray spot intensities and a power law in the tails. Furthermore, in this article it is pointed out that the variance of log spot intensity shows a positive correlation with the number of genes considered. Namely, the variance increases with the length of the arrays. In Purdom & Holmes (2005), the gene expression distribution is fitted using the Asymmetric Laplace distribution. The improvement upon the Gaussian distribution is notable. The Asymmetric Laplace presents asymmetry and heavy tails. One justification for the use of this distribution is based on the fact that it can be represented as the log-ratio of two independent random variables with Pareto distribution. Bhowmick *et al.* (2006) presents a statistical model for estimating gene expression using data from multiple laser scans is presented. These authors also point out that the distribution of gene expression exhibits heavy tails. A Cauchy distribution is adopted to model it.

In this work, we propose to model the gene expression distribution with an α-stable distribution. This distribution has been applied before to biology and physiology Zolotarev (1986); West & Deering (1994). However, as far as we know, this distribution has not been used before in cDNA dual dye microarray

data. We demonstrate that this distribution can very accurately fit the error distribution for gene expression. Furthermore, the α-stable distribution has many advantages when compared to other existing approaches in the literature, as will be emphasized in the paper.

A statistic to assess differential expression using an α-stable mixture model is presented. This statistic is based on the Scale Mixture of Normals property. The performance of the proposed method is compared to that of Lonnstedt & Speed (2002).

This paper is organized as follows: In Section 2, the α-stable distribution and its main properties are presented. In Section 3, we model the arrays from four different datasets with an α-stable distribution. In Section 4, the motivation and comparison of the proposed methodology to other existing approaches in the literature is discussed. Section 5 presents a statistic to assess differential expression using the properties of the α-stable distribution. In Section 6, the performance of the statistic proposed in this paper is tested. In Section 7, a possible application of the α-stable in the normalization of gene expression is proposed. Lastly, in Section 8, we summarize the conclusions of this work.

2 An overview of the α-stable distribution

The α-stable distribution is a family of distributions that presents heavy tails and is also capable of exhibiting a certain degree of asymmetry. This distribution has been used in the literature successfully to model skewed and impulsive phenomena. Furthermore, the α-stable distribution is a generalisation of the Gaussian distribution and allows us to describe impulsive processes by means of a small number of parameters.

This distribution has been widely studied in the literature and its properties are very well understood. It satisfies the Generalized Central Limit theorem which states that the limit distribution of infinitely many i.i.d. random variables, possibly with infinite variance distribution, is a stable distribution Feller (1966). The α-stable distribution also satisfies the stability property which states that any linear combination of random variables with α-stable distribution is also α-stable. More information on the main properties of this distribution can be found in Samorodnitsky & Taqqu (1994).

The α-stable distribution has four parameters, the shape parameter $\alpha \in (0, 2]$ is the characteristic exponent which sets the level of impulsiveness. $\beta \in [-1, +1]$ is a skewness parameter, ($\beta = 0$, for symmetric distributions and $\beta = \pm 1$ for the positive/negative stable family respectively). $\gamma > 0$ is the

DOI: 10.2202/1557-4679.1120

dispersion, a scale parameter and $\mu \in [-\infty, +\infty]$ is a shift parameter called location parameter.

There is no general closed expression for the α-stable probability density function (pdf); so that it is usually defined by its characteristic function, which is given by:

$$\varphi(\omega) = \begin{cases} e^{-|\gamma\omega|^{\alpha}[1-i\mathrm{sign}(\omega)\beta\tan(\frac{\pi\alpha}{2})]+i\mu\omega}, (\alpha \neq 1) \\ e^{-|\gamma\omega|[1+i\frac{2}{\pi}\mathrm{sign}(\omega)\beta\log(|\omega|)]+i\mu\omega}, (\alpha = 1) \end{cases} \quad (1)$$

Only for three particular cases is it possible to write the α-stable pdf. A distribution with characteristic exponent $\alpha = 2$ corresponds to a Gaussian distribution with $\gamma = \sigma/\sqrt{2}$ where σ is the standard deviation. The $\alpha = 1$ and $\beta = 0$ case corresponds to a Cauchy distribution and for $\alpha = 1/2$ and $\beta = 1$ to a Pearson distribution. Thus, the α-stable distribution can be seen as a generalization of the Normal distribution, and some features of linear system theory developed for Gaussian distribution can be applied directly to the α-stable distribution.

The α-stable density, except for the three particular cases mentioned above, must be calculated numerically. Moreover, it exhibits heavier tails than does a Gaussian distribution. In other words, it is more likely to obtain samples far from the mean for i.i.d. as an α-stable distribution with characteristic exponent $\alpha < 2$ than for the Gaussian case. This impulsive behaviour is a very well-known feature of the distribution of gene expressions.

When $\alpha < 2$, the tails probability $\{P < -\lambda\}$ and $\{P > \lambda\}$ as $\lambda \to \infty$, behave like the power law $\lambda^{-\alpha}$. This is also a known property of the distribution of gene expressions: the tails of the error distribution for gene expression data is well described by a power law (Paretian tail behaviour).

Let X be a vector with α-stable distribution and $0 < \alpha < 2$. Then,

$$E|X|^p < \infty \text{ for any } 0 < p < \alpha, \quad (2)$$

$$E|X|^p = \infty \text{ for any } p \geq \alpha. \quad (3)$$

Thus, α-stable random variables with $\alpha < 2$ have an infinite second-order moment. The standard deviation for a given random variable with an α-stable distribution does not converge to a meaningful value and an increase in the standard deviation is observed as the α-stable random vector lengthens.

As a means of showing the behaviour of the α-stable pdf, the stable density for varying α with $\beta = 0$ and varying β with $\alpha = 1.5$ are plotted in Figure 1. On the one hand, Figure 1a shows how the α parameter governs the degree of impulsiveness. Lower values of this parameter means heavier tails and higher

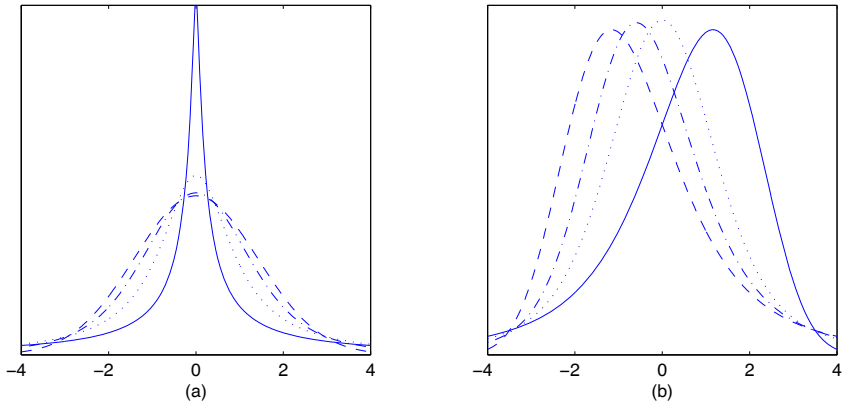

Figure 1: Density plot of α-stable distribution with location parameter $\mu = 0$ and $\gamma = 1$. (a) $\beta = 0$. Solid line: $\alpha = 0.5$. Dotted line: $\alpha = 1$. Dash-dotted line: $\alpha = 1.5$. Dashed line: $\alpha = 2$. (b) $\alpha = 1.5$. Solid line: $\beta = -1$. Dotted line: $\beta = 0$. Dash-dotted line: $\beta = 0.5$. Dashed line: $\beta = 1$.

peak of the α-stable distribution. On the other hand, Figure 1b shows an α-stable distribution with $\alpha = 1.5$ and varying β.

We fit data generated from an α-stable distribution with 5 different characteristic exponents using both Gaussian and α-stable distributions to show the performance difference between the two distributions. The characteristic exponents have been chosen to be possible values in real microarray experiments, as it will be shown in Section 3. The distribution is depicted in Figure 2. The shape of the Gaussian distribution (in dotted line) can change markedly with the number of samples because the variance is not defined for α-stable random variables.

3 Microarray data analysis

We model the distribution of gene expressions using the α-stable distribution for 4 different cDNA dual dye microarray datasets. The first dataset (labelled as 'self-self') consists of self-self hybridization of 19 different human cancer cell lines, the Stratagene universal reference RNA and RNA isolated from a tumor specimen Yang *et al.* (2002a). The second dataset ('zebrafish') are two sets of dye-swap experiments for a total of four replicate hybridizations. For each of these hybridizations, target cDNA from the swirl mutant was labelled using one of the Cy3 or Cy5 dyes, and the target cDNA wildtype mutant was labelled using the other dye. This experiment was carried out using zebrafish

DOI: 10.2202/1557-4679.1120

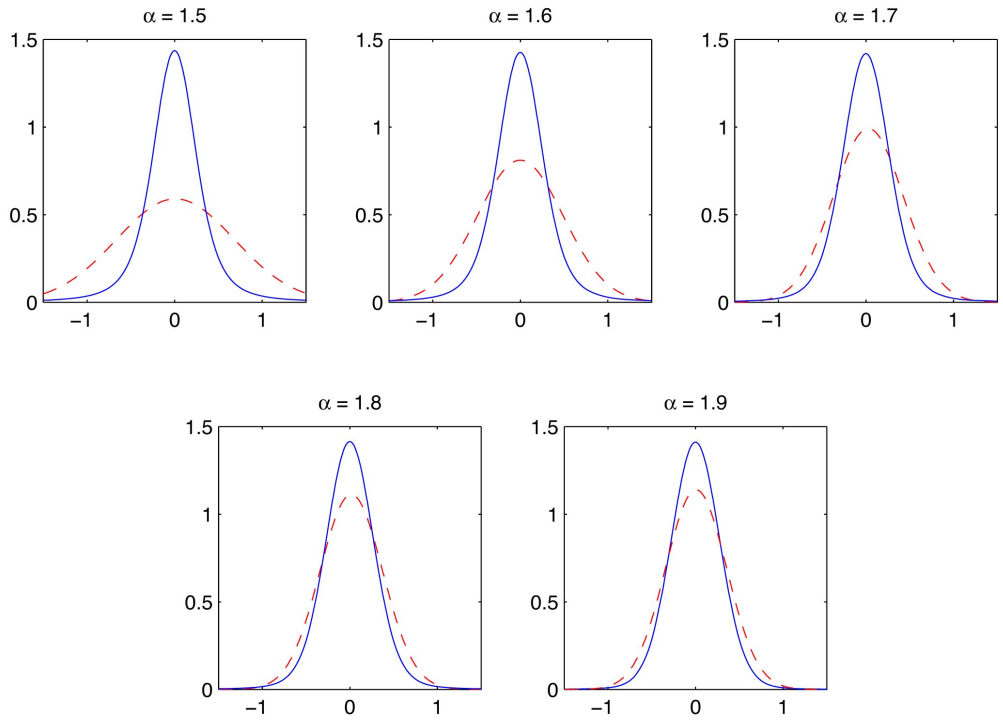

Figure 2: Continuous line: α-stable distribution. Dotted line: Gaussian fit.

as a model organism to study early development in vertebrates. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis[1]. The third dataset ('lymphoma') consists of tumor samples from diffuse large B-cell lymphoma patients Alizadeh *et al.* (2000). The last dataset ('yeast') is an analysis of regulatory variation in a cross between laboratory and wild strains of *Saccharomyces Cerevisiae* Yvert *et al.* (2003). These four datasets were chosen because they were also analyzed in Purdom & Holmes (2005), therefore it is possible to compare their results with our proposed methodology. Every dataset was normalized using locally weighted linear regression (LOWESS) Cleveland & Delvin (1988). This method is capable of removing intensity dependence in $\log_2(R_i/G_i)$ values and it has been successfully applied to microarray data Yang *et al.* (2002b). After normalization, each distribution of the gene expression has a similar shape: it exhibits heavier tails compared to Gaussian distribution and a certain degree of asymmetry.

There are different approaches to estimate the α-stable parameters Ku-

[1]This data is available as a dataset with the R package marrayClasses.

ruoglu (2001); Kogon & Williams (1998). For every array in the dataset, we estimate them using the maximum likelihood approach Nolan (2001). The parameter estimates are shown in Figure 3. It can be seen that the difference between the location (μ) and dispersion (γ) parameters estimated for the 'self-self' data is very low. For this dataset, the same RNA sample is labelled separately with green and red fluorescent dyes and hybridized to the same microarray; therefore, the gene expression distribution is expected to be symmetric. We find values of the skewness parameter β very close to zero in almost every case. There are only three cases in which β parameters are not near zero. They are the arrays 8, 18 and 24, (note that they are plotted with large circles in the figure). These three values are $\{\alpha_8 = 1.83, \beta_8 = -0.48\}$, $\{\alpha_{18} = 1.94, \beta_{18} = -0.65\}$ and $\{\alpha_{24} = 1.86, \beta_{24} = -0.77\}$, so that the α parameter for each of them is very close to 2. A well-known property of the α-stable distribution is that as the exponent α tends to the limiting value 2, more symmetric the α-stable distribution becomes and the less β parameter affects the shape. Therefore, these values of β are consistent with the expected symmetry of the distribution.

Figure 4 shows the distribution of the gene expression for an example array of each dataset. It can be seen that α-stable distribution fits the discrete distribution of the gene expression very accurately and better than does the Asymmetric Laplace or Gaussian. It is also seen that, despite the heavy tails and skewness of the Asymmetric Laplace distribution, this distribution has a very thin peak which is not always fit in gene expression data. Note how in the figure, the fit of the 'self-self' array considered is worse than the 'yeast' array using the Asymmetric Laplace distribution. The α-stable distribution, however, presents a smoother behaviour in the peak, which allows a better fit of the data. It can also be seen that the Gaussian distribution is not able to fit the gene expression data as the discrete histograms present heavier tails than the Normal distribution.

To compare numerically how α-stable, Asymmetric Laplace and the Normal distribution fit the gene expression distributions, we calculated the Kullback-Leibler, χ^2 and Hellinger distance Borovkov (1998). The χ^2 distance penalizes possible outliers in the fitting. Namely, a small amount of samples affects the measured χ^2 distance more than the Hellinger and K-L distance. The former is the most robust to outliers among the three distances considered. These tests were applied to each array and better performance for the α-stable distribution was obtained for all of them. Table 1 shows the corresponding mean and standard deviation of these tests for each dataset. It is shown that the α-stable distribution fits the empirical gene expression distribution much better than does the Asymmetric Laplace or Gaussian. Furthermore,

DOI: 10.2202/1557-4679.1120

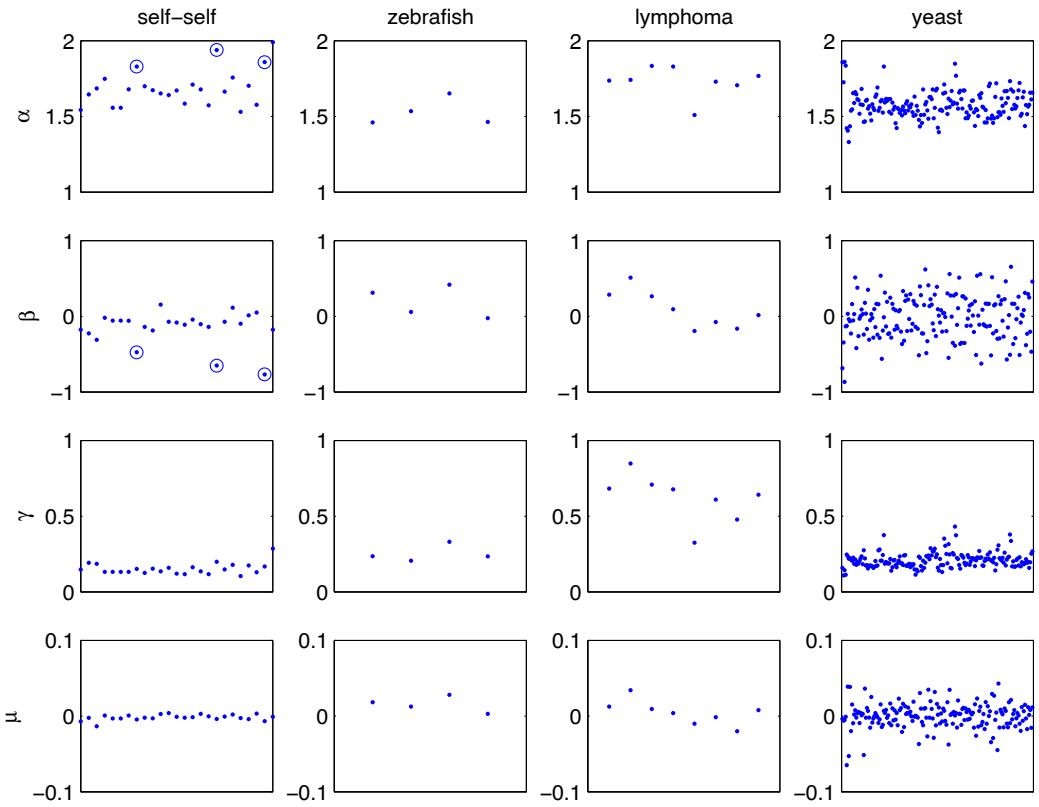

Figure 3: Estimated parameters for each dataset. First row: characteristic exponent α. Second row: skewness parameter β. Third row: dispersion γ. Fourth row: location parameter μ.

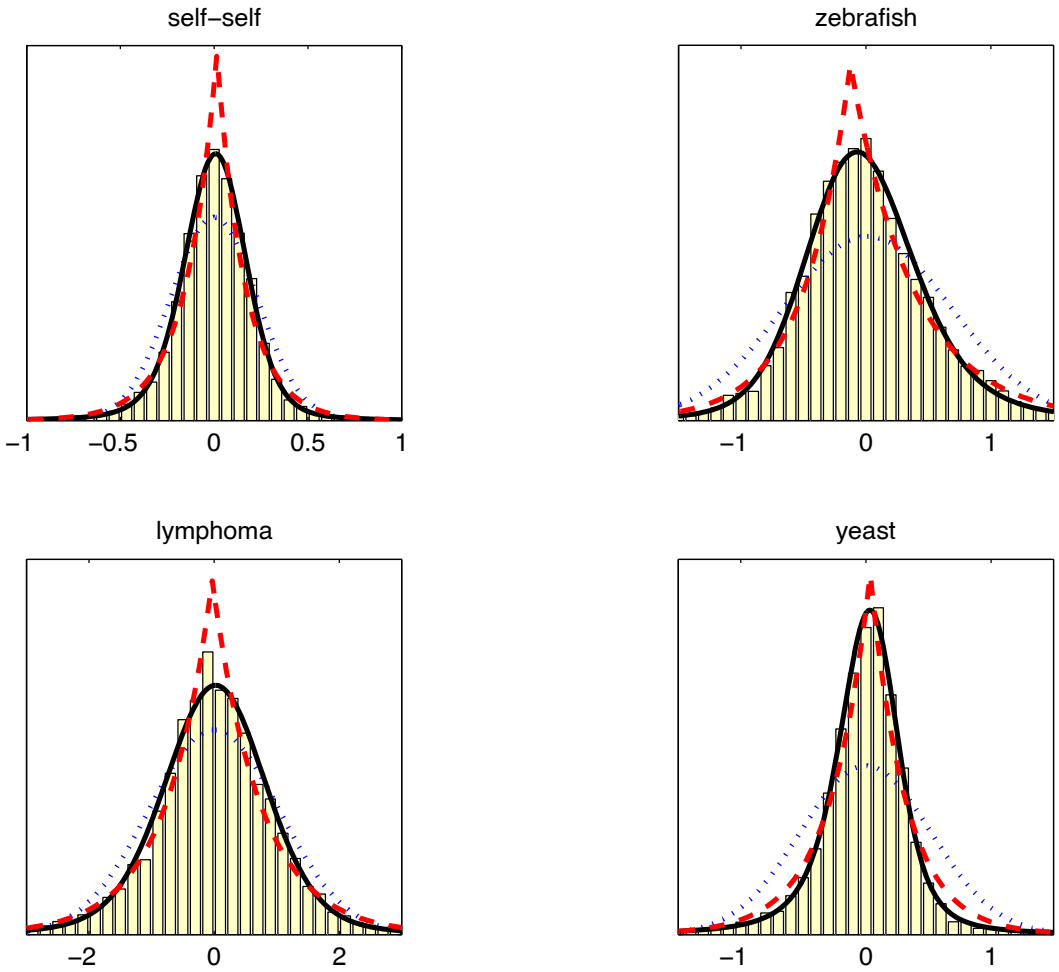

Figure 4: empirical gene expression histogram and predicted density for one array of each dataset. From the 'self-self' dataset we choose the array 9 (NT2.2(testis)). From the 'zebrafish' and 'lymphoma' dataset, the 2 array and DLCL-0024 are chosen respectively. From the 'yeast' dataset, we used 14-4-aCy3. Solid line: α-stable distribution. Dashed line: Asymmetric Laplace distribution. Dotted line: Gaussian distribution.

DOI: 10.2202/1557-4679.1120

Table 1: Kullback Leibler, χ^2 and Hellinger distance between the empirical gene expression distribution and the predicted stable, Asymmetric Laplace and Gaussian density for each dataset. The number denotes the mean of the distance calculated for each dataset. In brackets, the error (standard deviation). In bold, the lowest distance and standard deviation.

	self-self	zebrafish	lymphoma	yeast
KL(Stable)	**0.013 (0.005)**	**0.0171 (0.0020)**	**0.021 (0.003)**	**0.018 (0.005)**
KL(ALaplace)	0.022 (0.019)	0.066 (0.021)	0.022 (0.003)	0.047 (0.017)
KL(Gauss)	0.10 (0.04)	0.35(0.10)	0.07 (0.03)	0.25 (0.09)
χ^2(Stable)	**0.015 (0.008)**	**0.015 (0.003)**	**0.022 (0.004)**	**0.016 (0.005)**
χ^2(ALaplace)	0.04 (0.06)	0.074 (0.019)	0.038 (0.008)	0.058 (0.022)
χ^2(Gauss)	0.12 (0.05)	0.6 (0.3)	0.09(0.05)	0.39 (0.21)
Hell.(Stable)	**0.0036 (0.0021)**	**0.0043 (0.0007)**	**0.0061 (0.0011)**	**0.0044 (0.0014)**
Hell.(ALaplace)	0.009 (0.009)	0.020 (0.005)	0.0087 (0.0015)	0.015 (0.005)
Hell.(Gauss)	0.030 (0.012)	0.11 (0.04)	0.025 (0.012)	0.07 (0.03)

the standard deviation is considerably lower for the α-stable case. This means that the Asymmetric Laplace and Gaussian, contrary to α-stable, fits the gene expression distribution accurately or poorly depending on the array. This fact was remarked in the last paragraph and illustrated in Figure 4. The lower mean and standard deviation calculated for the K-L, χ^2 and Hellinger distance for the α-stable distribution, shows that this distribution accurately fits for each array.

4 Comparison with previous work

- Kuznetsov (2001) noted that the gene expression distribution follows a Pareto-like distribution. He modelled the gene expression distribution using several classes of skewed probability functions and found better performance using the Pareto-like distribution. He introduced an artificial location parameter to generalize the Pareto distribution. However, we would like to point out that α-stable distribution already accounts for this parameter and provides a good fit in both the main lobe and the tails of the distribution. The above author demonstrated that the empirical histograms of gene expression levels are well fitted by a power-law distribution. The α-stable distribution also has a Paretian tail behaviour when $\alpha < 2$. Specifically, if X is a random variable with α-stable distribution with $\alpha < 2$, then Samorodnitsky & Taqqu (1994):

$$\lim_{\lambda \to -\infty} \lambda^{\alpha} \{P < \lambda\} = C_{\alpha} \frac{1+\beta}{2} \gamma^{\alpha} \tag{4}$$

$$\lim_{\lambda \to -\infty} \lambda^{\alpha} \{P < -\lambda\} = C_{\alpha} \frac{1-\beta}{2} \gamma^{\alpha} \tag{5}$$

where

$$C_{\alpha} = \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} \text{ if } \alpha \neq 1 \tag{6}$$

$$C_{\alpha} = \frac{2}{\pi} \text{ if } \alpha = 1 \tag{7}$$

Furthermore, Mandelbrot remarked the fact that the use of the Stable distribution for describing empirical principles was preferable to the use of Zipf-Pareto distribution for both, theoretical and practical reasons. (See Zolotarev (1986) for a deeper explanation regarding Stable laws in biology).

- In Hoyle *et al.* (2002), a wide range of datasets are analyzed. The error distribution is approximated by a log-normal in the bulk of microarray spot intensities which is claimed to be a good approximation for the distribution of most of the spot-intensity values. It is also pointed out that the tails of the distribution agree well with Zipf's law, a special case of Pareto behaviour (or power law) Newman (2005). Therefore, two different distributions are used to model the distribution of gene expression, log-normal in the bulk and power law in the tails. Two possible, and heuristic, explanations for this different behaviour are given in Hoyle *et al.* (2002). Contrary to this, the α-stable distribution enables us to model the gene expression distribution with only one distribution, modelling very accurately both the centre and the tails.

  Furthermore, in Hoyle *et al.* (2002), it is pointed out that the variance σ^2 of log spot intensity increases as the number of genes considered increases. This result agrees well with the properties of the α-stable distribution, as stated in Section 2. The variance is not defined for stable processes with $\alpha < 2$, therefore the second-order statistics cannot help us to gain an insight into stable random variables. Due to this fact, the standard deviation of a random variable with α-stable distribution increases as the length of the random vector increases and it does not converge to a given value.

- In Purdom & Holmes (2005), an Asymmetric Laplace distribution is used to fit the gene expression distribution and its performance is compared to the Gaussian distribution. In the previous section, our methodology was compared experimentally to the Asymmetric Laplace distribution and it was shown that this distribution does not always properly fit the

DOI: 10.2202/1557-4679.1120

gene expression distribution. The histogram of gene expression levels often presents a smoother behaviour in the maximum. Furthermore, although the Asymmetric Laplace distribution presents heavier tails than the Gaussian distribution, the tails of this distribution are exponential, not algebraic, and do not exhibit Paretian behaviour. A Laplace and Asymmetric Laplace distribution for identification of differential expression in microarray experiments have been assumed recently in Bhowmick *et al.* (2006). We believe that an α-stable assumption for the gene expression distribution could help in building new statistical methods to assess whether a gene is differentially expressed or not.

- In Khondoker *et al.* (2006), the distribution of gene expression is modelled using a Cauchy distribution as a part of a statistical model for estimating gene expression using data from multiple-laser scans. The Cauchy distribution is chosen rather than assuming a Normal distribution in order to take into account the outliers. In our work, we assume neither Gaussian nor Cauchy, but both are particular cases of the α-stable family. Specifically, for $\alpha = 1$, $\beta = 0$, the inverse Fourier transform of the characteristic function in Eq. 1 has an analytical solution. In that particular case, the pdf of the α-stable is

$$\frac{\gamma}{\pi((x-\mu)^2+\gamma^2)} \tag{8}$$

  which corresponds to a Cauchy distribution with location parameter μ and dispersion parameter γ. If the distribution of gene expression were Cauchy, we would have found that $\alpha \approx 1$ and $\beta \approx 0$ most of the times for the characteristic exponent and the skewness parameter, respectively; and Figure 3 shows that the values reached in the estimation of the shape parameter α are typically in the interval $[1.5 - 1.8]$.

The α-stable distributions provide a unified framework for modelling various characteristics that were modelled by other models in an isolated way; hence this evidence suggests that we could use α-stable distribution to model the distribution of the gene expression.

5 Assessing differential expression using the α-stable

The Gaussian distribution is a particular case of the α-stable distribution. Therefore, the α-stable assumption is a long-tailed and skewed alternative

that offers the same results in identification of differential expression, if the distribution of the gene expression were Gaussian, as those based on the Normal distribution Lonnstedt & Speed (2002). However, the distribution of the gene expression is found empirically not to be Gaussian.

There are some works in the literature which, following Purdom & Holmes (2005), assume a Laplace distribution for gene error expression. In Bhowmick *et al.* (2006), a Laplace mixture model is proposed as a long-tailed alternative to the Normal distribution in order to identify differentially expressed genes in microarray experiments.

The independence assumption is frequently made between genes, but in the presence of differential expression, assuming an identical value of the μ parameter for all genes is not a realistic scenario. For that reason, many works consider that the log-ratios of non-differentially expressed genes are distributed around zero ($\mu = 0$) and they propose a two-component mixture model differing in the location parameter Lonnstedt & Speed (2002); Bhowmick *et al.* (2006); Lewin *et al.* (2007). In particular, Lonnstedt & Speed (2002) assumes the following model:

$$M_{ij}|\mu_i, \lambda_i, \sigma \sim N(\mu_i, \lambda_i\sigma^2) \text{ for all } i. \tag{9}$$

where N is the number of genes on each array and n the number of replicates (arrays), and the data will be denoted as $M_{ij} = log\left(\frac{R_{ij}}{G_{ij}}\right)$, where $i = 1...N$, $j = 1...n$ and λ follows an Inverse Gamma distribution. This is an Scale Mixture of Normals model, although it is not explicitly stated in Lonnstedt & Speed (2002) and, therefore, if $\mu_i = 0$, M_{ij} is distributed as a t-student (see Fernandez & Steel (2000)).

Moreover, the parameter μ is regarded as drawn from a distribution $P(\theta)$ if the gene is expressed or $\mu = 0$ if it is not differently expressed.

$$p(\mu_i|\lambda_i, \sigma) = wP(\theta) + (1 - w)\delta(0) \tag{10}$$

A similar model could be build using the Scale Mixture of Normals property of the α-stable distribution. If the prior distribution for λ in equation (9) is a positive α-stable, instead of Inverse Gamma, with the following values of the parameters

$$p(\lambda_i) = f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}}, 0) \tag{11}$$

then the distribution of M_{ij} for the non-expressed genes is distributed as a symmetric α-stable, which was found to be the case when the self-self dataset was studied. Then, some well-known properties of the α-stable distribution can

DOI: 10.2202/1557-4679.1120

be used to build a statistic S to assess whether a gene is differently expressed, as will be explained below.

Let z_i indicate whether a given gene is differentially expressed ($z_i = 1$) or not ($z_i = 0$):

$$z_i = \begin{cases} 0 & \text{if } \mu_i = 0, \\ 1 & \text{if } \mu_i \sim f_{\alpha,0}(\sigma, 0). \end{cases}$$

The log posterior ratio for a given gene i can be calculated as

$$S_i = \log \frac{Pr(z_i = 1|M_{ij})}{Pr(z_i = 0|M_{ij})}. \tag{12}$$

Following the Bayes' Theorem and assuming independence between genes

$$S_i = \log \frac{w}{1-w} \frac{Pr(M_i|z_i = 1)}{Pr(M_i|z_i = 0)}, \tag{13}$$

where M_i is the vector of the n replicates for gene i. Our goal is to calculate the posterior probabilities $Pr(M_i|z_i = 1)$ and $Pr(M_i|z_i = 0)$ in order to compute the log posterior odds S_i which, for a given gene i, computes the probability of being differently expressed.

In parallel to the work of Lonnstedt & Speed (2002), the statistic S_i can be considered a way of ranking genes. Therefore, the proportion of expressed genes needs to be established *a priori*.

Considering $M_{i\cdot}$ as the average of M_{ij} for $j = 1 \dots n$ for a given gene i. The distribution of M_i conditional on μ_i, λ_i and σ is

$$\begin{aligned} p(M_i|\mu_i, \lambda_i, \sigma) &= (2\pi\lambda_i)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j (M_{ij}-\mu_i)^2} \\ &= (2\pi\lambda_i)^{-\frac{n}{2}} \sigma^{-n} e^{-\frac{1}{2\lambda_i\sigma^2}(\sum_j (M_{ij}-M_{i\cdot})^2 + n(M_{i\cdot}-\mu_i)^2)}, \end{aligned} \tag{14}$$

and, for the proposed model:

$$\begin{aligned} p(M_i|z_i = 1) &= \int\int p(M_i, \mu_i, \lambda_i) d\mu_i d\lambda_i \\ &= \int\int p(M_i|\mu_i, \lambda_i) p(\mu_i|\lambda_i, z_i = 1) p(\lambda_i) d\mu_i d\lambda_i \end{aligned} \tag{15}$$

and

$$\begin{aligned} p(M_i|z_i = 0) &= \int\int p(M_i|\mu_i, \lambda_i) p(\mu_i|\lambda_i, z_i = 0) p(\lambda_i) d\mu_i d\lambda_i \\ &= \int p(M_i|\lambda_i) p(\lambda_i) d\lambda_i. \end{aligned} \tag{16}$$

Substituting in the expressions (15) and (16) the distributions for our model and considering the following equality

$$N(\mu_i|0,\lambda_i\sigma^2)\cdot e^{-\frac{n(M_{i.}-\mu_i)^2}{2\lambda_i\sigma^2}} = N(\mu_i|\frac{n}{n+1}M_{i.},\frac{\lambda_i\sigma^2}{n+1})\cdot(n+1)^{-1/2}\cdot e^{-\frac{1}{2\lambda_i\sigma^2}\frac{n}{n+1}M_{i.}^2}, \tag{17}$$

it is possible to integrate out the distribution $N(\mu_i|\frac{n}{n+1}M_{i.},\frac{\lambda_i\sigma^2}{n+1})$ and to obtain the following integrals:

$$\begin{aligned} p(M_i|z_i=1) &= \int(2\pi\lambda_i)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j(M_{ij}-M_{i.})^2} \\ &\times (n+1)^{-1/2}e^{-\frac{1}{2\lambda_i\sigma^2}\frac{n}{n+1}M_{i.}^2} \\ &\times f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}},0)d\lambda_i \end{aligned} \tag{18}$$

and

$$\begin{aligned} p(M_i|z_i=0) &= \int(2\pi\lambda_i)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i\sigma^2}\sum_j(M_{ij})^2} \\ &\times f_{\frac{\alpha}{2},1}(2\{\cos\left(\frac{\pi\alpha}{4}\right)\}^{\frac{2}{\alpha}},0)d\lambda_i. \end{aligned} \tag{19}$$

Due to the non existence of an analytical expression for the α-stable pdf, the integrals (18) and (19) need to be calculated numerically. Some different approaches could be used in order to accomplish this goal. We took advantage of the fact that drawing samples from an α-stable distribution can be easily accomplished using Chambers' algorithm Chambers *et al.* (1976). If we have T random samples $\left[\lambda_i^{(1)}...\lambda_i^{(t)}...\lambda_i^{(T)}\right]$ with distribution $p(\lambda_i)$ given by Eq. (11), a Monte Carlo empirical estimate of the integrals (18) and (19) is

$$\begin{aligned} p(M_i|z_i=1) &= \frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j(M_{ij}-M_{i.})^2} \\ &\times (n+1)^{-1/2}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\frac{n}{n+1}M_{i.}^2} \end{aligned} \tag{20}$$

and

$$p(M_i|z_i=0) = \frac{1}{T}\sum_{t=1}^{T}(2\pi\lambda_i^{(t)})^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\lambda_i^{(t)}\sigma^2}\sum_j(M_{ij})^2}. \tag{21}$$

DOI: 10.2202/1557-4679.1120

6 Results

6.1 Case study 1

To illustrate the performance of the statistic S proposed in this paper, we simulated a dataset containing $N = 10,000$ genes and $n = 4$ replicates. The non-expressed genes were simulated following an α-stable distribution with parameters: $\alpha = 1.8$, $\beta = 0$, $\sigma = 0.1$ and $\mu = 0$. Thus, the samples are simulated from the following mathematical model:

$$M_{ij}|\mu_i, \lambda_i, \sigma \sim N(0, \lambda_i 0.1^2) \text{ for } i = 1 : N. \tag{22}$$

$$p(\lambda_i) = f_{\frac{1.8}{2},1}(2\{\cos\left(\frac{1.8\pi}{4}\right)\}^{\frac{2}{1.8}}, 0). \tag{23}$$

They were typical values obtained in the analysis of the four gene expression datasets studied in Section 3.

A proportion $p = 0.01$ of the $N = 10,000$ genes were considered to be differently expressed. For that set of genes, the values of the α-stable parameters were chosen the same as the non-expressed but the location parameter μ was simulated as an α-stable with dispersion parameter set to $V\sigma$ with $V = 1.5$, where V represents a type of generalized signal-to-noise ratio. The parameter V was also introduced in Lonnstedt & Speed (2002); Bhowmick *et al.* (2006). The estimation of this parameter is very difficult because only a very small proportion of genes are expressed and we do not know which ones. This difficulty was also pointed out in Lonnstedt & Speed (2002) for a Gaussian mixture model and for a Laplace mixture model in Bhowmick *et al.* (2006) where the parameter V was not estimated correctly. It will be shown in the simulation study that the performance of our algorithm is not affected by the ignorance of this parameter. This is for two different reasons: on the one hand, the α-stable distribution is a heavy-tailed distribution, and therefore it is a proper distribution to accommodate outliers in the data; on the other hand, the number of genes differently expressed is usually a very small proportion of the whole dataset.

One of the simulated datasets is plotted in Figure 5, where the average M-values versus the logarithm of the variance is plotted. The expressed genes are denoted with crosses. These are the genes with expectation different to zero. It is easily seen that many of the true influenced genes have a negligible value of the average, and therefore it is not possible to detect them using any method.

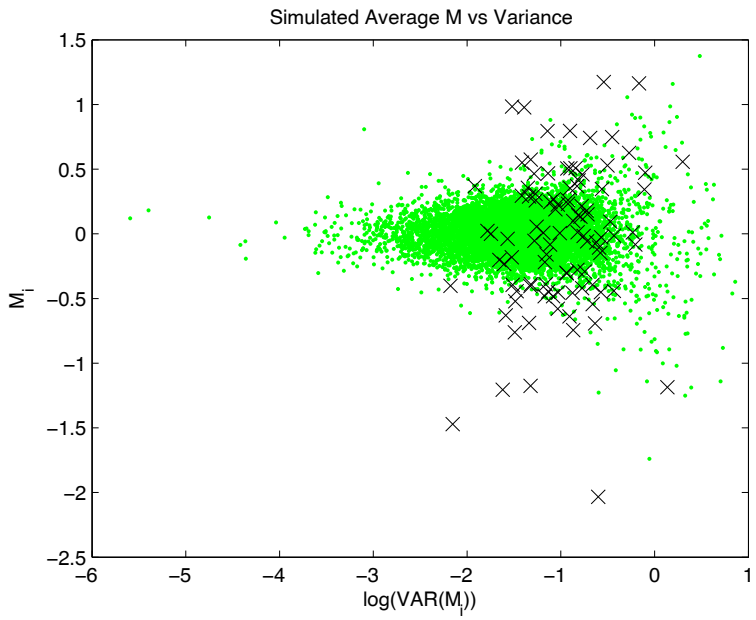

Figure 5: $M_{i.}$ vs. log-variance for one of the simulated datasets. *Crosses:* True expressed genes.

For each different values of the cutoff w considered (40 different values from $w = 0$ to $w = 1$), 100 different datasets were simulated. The Stable statistic S and B were calculated for each dataset.

The Receiver Operating Characteristic curve for the 40 different cutoffs w is plotted in Figure 6 for each synthetic dataset. The fraction of true-positive and false-positive genes is averaged over the 100 datasets. In this figure, the statistic based on the α-stable distribution is compared with B, the statistic based on the scale t-statistics proposed in Lonnstedt & Speed (2002). The Stable statistic exhibits higher values of true positives and true negatives than B for each value of the cutoff.

6.2 Case study 2

In the case study 1, we compared the proposed method with a published method using the receiver operating characteristic curve. In that case, the data was simulated using the α-stable mixture model; therefore, the proposed methodology was expected to perform better. The use of the α-stable mixture model instead of the Gaussian mixture to model the distribution of gene expression was justified in Section 3 and 4. Although the appropriateness of

DOI: 10.2202/1557-4679.1120

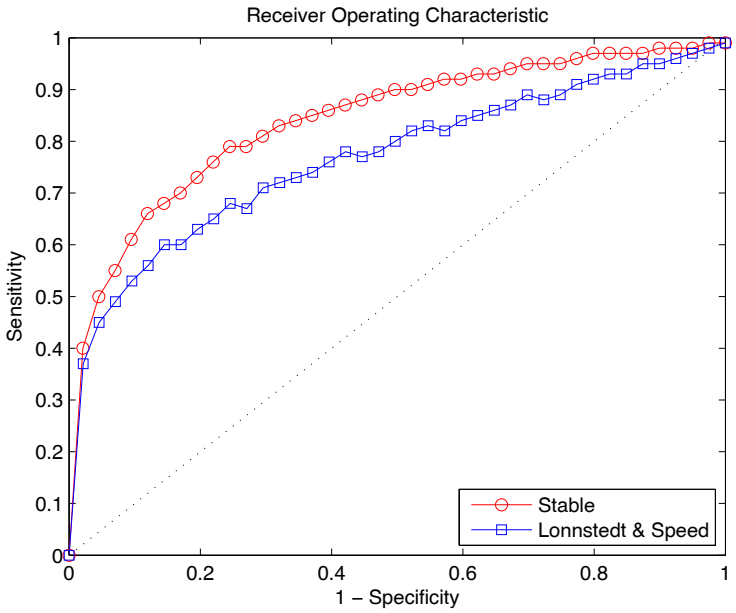

Figure 6: Case Study 1. Receiver Operating Characteristic curve for S_i and B_i statistics computed on the simulated α-stable data with $\alpha = 1.8$, $\beta = 0$, $\sigma = 0.1$ and $\mu = 0$.

the α-stable modelling has been shown in this article, in the current section the data will be generated using the Gaussian mixture model in Lonnstedt & Speed (2002). The main difference between the two models is the expression of the prior distribution $p(\lambda)$. For ν degrees of freedom and scale parameters $a > 0$, $c > 0$, Lonnstedt & Speed (2002) set $\tau_i = na/2\sigma_i^2$ and assume that

$$\tau_i \sim \Gamma(\nu, 1) \tag{24}$$

$$\mu_i|\tau_i = \begin{cases} 0 & \text{if } z_i = 0, \\ N(0, \frac{cna}{2\tau_i}) & \text{if } z_i = 1. \end{cases}$$

In this simulation, the following values of the parameters were chosen: $N = 10000$ genes, $n = 4$ replicates, $\nu = 2.8$, $a = 0.040$ and $c = 1.5$. The proportion of expressed genes was set to $p = 0.01$. These hyperparameters were estimated from a real dataset (see Lonnstedt & Speed (2002) for a deeper explanation of the model and the hyperparameters).

For each different value of the cutoff w (12 different values from $w = 0$ to $w = 1$), 100 different datasets were simulated. The Stable statistic S and B

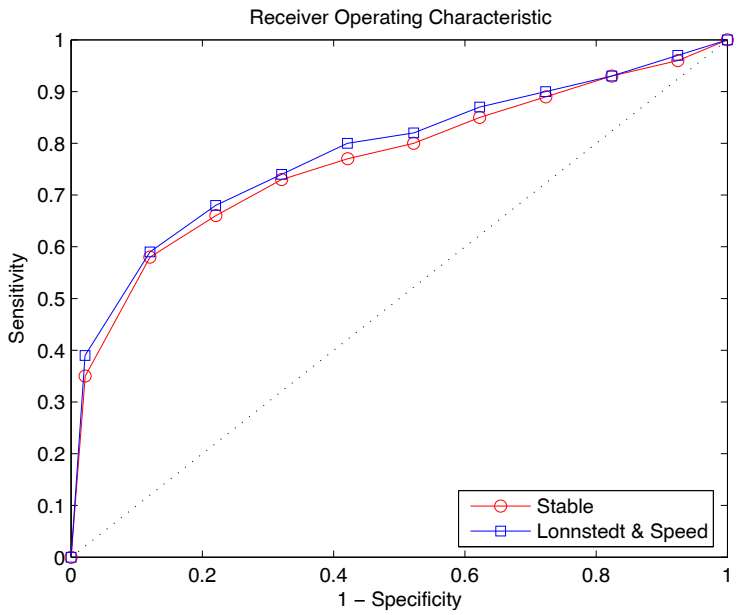

Figure 7: Case Study 2. Receiver Operating Characteristic curve for S_i and B_i statistics computed on the data simulated using the Gaussian mixture model in Lonnstedt & Speed (2002).

were calculated. In Figure 7, both statistics are compared. The B statistic performs only slightly better than S for this dataset, but it is important to note that, even for this dataset, the B statistic does not improve very much upon the stable statistic. This was expected since α-stable is a scale mixture of Gaussians. Furthermore, the α-stable distribution is more flexible than the Gaussian distribution, because it has a greater number of parameters. This feature allows the α-stable distribution to fit a large family of distributions and illustrates that the proposed methodology is a useful alternative to assess whether a gene is differently expressed.

7 Future work: normalization of gene expression

Some normalization methods in the literature assume that the fluctuations across replicated microarray data follow a Gaussian distribution Chen *et al.* (1997); Wang *et al.* (2002). However, the fluctuations across replicates is found to present a degree of impulsiveness which cannot be modelled using a Gaussian

DOI: 10.2202/1557-4679.1120

distribution. Consequently, there are some recent works in the literature which, following Purdom & Holmes (2005), assume a Laplace distribution for gene error expression. In Ramirez *et al.* (2006), a new method for normalization of cDNA Microarray data is proposed based on the Laplace distribution and its properties.

Our work suggests that using the α-stable distribution and its well-studied properties instead of a Laplace distribution could help to build a novel method of normalization. We believe that the α-stable distribution is a skewed alternative which could enable us to build not only new statistics to assess whether differential expression has occurred, as it has been pointed out in the previous section, but also to develop novel normalization algorithms. Furthermore, the properties of the α-stable distribution are very well understood and, to our knowledge, they have not been applied before in microarray data.

8 Conclusion

In this work, we have presented a new statistical model for the distribution of differential gene expression. The model provides the flexibility for modelling impulsiveness and skewness required for gene expression data. We stress the fact that it is not an *ad-hoc* model but has strong theoretical justifications such as the generalised central limit theorem. It confirms with earlier observations made by other researchers such as Paretian tails and non-converging standard deviation. Both impulsiveness and skewness are parametrised in a parsimonious way using the α-stable distribution. A rich variety of techniques exist in the literature for parameter estimation.

A statistic based on α-stable modelling to assess differential expression in replicated microarray data is presented. A mixture of an α-stable and a Dirac delta function is introduced to model the expressed and non-expressed genes, respectively. The Scale Mixture of Normals property is used to calculate the Bayes log posterior odds. The performance was compared to a statistic based on t-student distribution. We believe that the statistical model presented in this paper will be very useful in estimation and detection problems involving gene expression array data.

References

Alizadeh, Ash A., Eisen, Michael B., Davis, R. Eric, Ma, Chi, Lossos, Izidore S., Rosenwald, Andreas, Boldrick, Jennifer C., Sabet, Hajeer, Tran, Truc, Yu, Xin, Powell, John I., Yang, Liming, Marti, Gerald E., Moore, Troy, James Hudson, Jr, Lu, Lisheng, Lewis, David B., Tibshirani, Robert, Sherlock, Gavin, Chan, Wing C., Greiner, Timothy C., Weisenburger, Dennis D., Armitage, James O., Warnke, Roger, Levy, Ronald, Wilson, Wyndham, Grever, Michael R., Byrd, John C., Botstein, David, Brown, Patrick O., , & Staudt, Louis M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Bhowmick, Debjani, Davison, A. C., Goldstein, Darlene R., & Ruffieux, Yann. 2006. A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics*, **7**(4), 630–641.

Borovkov, A. A. 1998. *Mathematical statistics*. Amsterdam: Gordon and Breach science.

Chambers, J., Mallows, C., & Stuck, B. 1976. A method for simulating stable random variables. *Journal of the American Statistical Association*, **71**, 340–344.

Chen, Y., Doucherty, E., & M., Bittner. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, **2**, 364–374.

Cleveland, W. S., & Delvin, S. J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of American Statistical Association*, **83**(403), 596–610.

Feller, W. 1966. *An introduction to probability theory and its applications*. Vol. II. New York: Wiley & Sons.

Fernandez, Carmen, & Steel, Mark F. J. 2000. Bayesian regression analysis with scale mixture of Normals. *Econometric Theory*, **16**, 80–101.

Gottardo, Raphael, Panucci, James A., Kuske, Cheryl R., & Brettin, Thomas. 2003. Statistical analysis of microarray data: a Bayesian approach. *Biostatistics*, **4**(4), 597–620.

Hoyle, D. C., Rattray, M., Jupp, R., & Brass, A. 2002. Making sense of microarray data distributions. *Bioinformatics*, **18**(4), 576–584.

DOI: 10.2202/1557-4679.1120

Khondoker, Mizanur R., Glasbey, Chris A., & Worton, Bruce J. 2006. Statistical estimation of gene expression using multiple laser scans of microarrays. *Bioinformatics*, **22**(2), 215–219.

Kogon, S. M., & Williams, D. B. 1998. *Characteristic function based estimation of stable parameters.* Boston, MA: R. Feldman, and M. Taqqu (Eds.), A Practical Guide to Heavy Tailed Data.

Kuruoglu, E. E. 2001. Density Parameter Estimation of Skewed alpha-Stable Distributions. *IEEE Transactions on Signal Processing*, **49**(10), 2192–2201.

Kuznetsov, Vladimir A. 2001. Distribution Associated with Stochastic Processes of Gene Expression in a Single Eukaryotic Cell. *EURASIP Journal on applied signal processing*, **4**, 285–296.

Lewin, Alex, Bochkina, Natalia;, & Richardson, Sylvia. 2007. Fully Bayesian Mixture Model for Differential Gene Expression: Simulations and Model Checks. *Statistical Applications in Genetics and Molecular Biology*, **6**(1).

Lonnstedt, Ingrid, & Speed, Terry. 2002. Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, **46**, 323.

Nolan, J. P. 2001. Maximum likelihood estimation of stable parameters. *in Levy Processes (O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, eds.)*, 379–400.

Purdom, Elizabeth, & Holmes, Susan. 2005. Error distribution for gene expression data. *Statistical applications in genetics and molecular biology*, **4**(1).

Ramirez, J., Paredes, J.L., & Arce, G. 2006. Normalization of cDNA Microarray Data Based on Least Absolute Deviation Regression. *In: IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse, France.*

Samorodnitsky, G., & Taqqu, M.S. 1994. *Stable Non-Gaussian Random Process: Stochastic Models with Infinite Variance.* New York: Chapman-Hall.

Wang, Y., Lu, J., R., Lee., Gu, Z., & Clarke, R. 2002. Iterative normalization of cDNA microarray data. *IEEE Trans. Inform. Technol. Biomed.*, **6**, 29–37.

West, Bruce J., & Deering, William. 1994. Fractal Physiology for Physicists: Levy Statistics. *Physics Reports*, **246**(1-2), 1–100.

Yang, Ivana, Chen, Emily, Hasseman, Jeremy, Liang, Wei, Frank, Bryan, Wang, Shuibang, Sharov, Vasily, Saeed, Alexander, White, Joseph, Li, Jerry, Lee, Norman, Yeatman, Timothy, & Quackenbush, John. 2002a. Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biology*, **3**(11), research0062.1–research0062.12.

Yang, Yee Hwa, Dudoit, Sandrine, Luu, Percy, Lin, David M., Peng, Vivian, Ngai, John, & Speed, Terence P. 2002b. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids Res.*, **30**(4), e15–.

Yvert, Gael, Brem, Rachel B., Whittle, Jacqueline, Akey, Joshua M., Foss, Eric, Smith, Erin N., Mackelprang, Rachel, & Kruglyak, Leonid. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, **35**(1), 57–64.

Zolotarev, V. M. 1986. *One dimensional Stable Distributions*. Providence: Translation on Mathematical Monographs 65. American Math. Soc.

DOI: 10.2202/1557-4679.1120