CHAPTER 3

Application of Skew Slash Family of Distributions to Microarray Data

3.1 Introduction

Novel techniques for analyzing microarray data are constantly being developed and are usually tested against the known structure of normally distributed data. However, microarray data are not, in fact, normally distributed, and testing against such data can have misleading consequences. Microarray data are non-normally distributed under any of the standard data transformations. The resulting data tend to have heavier tails than normal, and they are often more skewed as well. Hence the standard methodologies for identifying differentially expressed genes can give unexpected and misleading results when the data are not normally distributed. Robust methods should be used when analyzing microarray data.

Results included in this chapter form the paper Bindu (2011b, 2012b, 2012c, 2012d and 2013a).
Microarray intensity data present two problems that can be solved by using transformations. The first one is the intensity dependencies, that usually appear in MA plots even when the data come from self-self experiments. The second one is the mean variance dependence, i.e., higher intensity measures show higher variability. To partially cope with these drawbacks, a standard microarray data analysis is carried out using the log$_2$ transformation of the intensity data (see Speed, 2003) together with normalization methods to eliminate structures on MA plots. This approach makes it possible to work with log-ratios, which is a simple and intuitive transformation for biologists.

In this Chapter we proposed the skew slash family of distributions developed in Chapter 2 as an approximation of the distribution of the log-ratios of measured gene expression across genes.

3.2 Applications of Skew Slash Distributions

In this section, we first illustrate the applications of the skew slash distributions: Skew slash distributions generated by Cauchy Kernal, Skew slash distributions generated by normal Kernal, Skew slash t and asymmetric slash Laplace. Then, we compared the goodness of fit of skew slash distributions ($SSCL$, $SSNL$, $SST$ and $ASL$) for microarray gene expression data.

3.2.1 Fitting Skew Slash Cauchy-Laplace to Two-Color Microarray Data

Here we discuss the applications of the Skew slash distributions generated by Cauchy Kernal. In this section we illustrate the applications of $SSCL$ for microarray gene expression data and we compared it with the normal and skew slash normal distributions. For this purpose we downloaded the cDNA dual dye microarray dataset (Experiment id-38067) from the Stanford Microarray Database. The microarray was performed on cDNA chips manufactured by Stanford Functional genomic facility. Each array chip contains approximately 42000 human cDNA elements, representing over 30000 unique genes. The dataset was normalized using Smoothing Scatterplots by Locally Weighted Regression (LOWESS) (Cleveland
and Delvin, 1988) and is given in Figure 3.1. 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., 2002), where $R_i$ is the red dye intensity and $G_i$ is the green dye intensity for the $i^{th}$ gene. The Figure 3.2 shows the box plots of intensities before and after normalization.

After normalization, each distribution of the gene expression has a similar shape and exhibits heavier tails compared to a Gaussian distribution and a certain degree of asymmetry. We estimated the parameters using maximum likelihood estimation. In our illustrations, the maximization of the likelihood is implemented using the \texttt{optim} function of the R software, applying the \texttt{BFGS} algorithm (R Development Core Team, 2006). Also one can use the function \texttt{nlminb} in the S-Plus package to locate the maximum point of the likelihood function assuming all the parameters are unknown.

We fitted the \textit{SSCL} to the normalized microarray intensities. The estimated MLEs and standard errors (SE) for the parameters are given in Table 3.1. Figure 3.3 depicts the histogram of the gene expression data with the fitted \textit{SSCL} along with skew slash probability density function. We compared the empirical distribution of the microarray gene expression and the \textit{SSCL} distribution. It can be clearly seen that the estimated density of the \textit{SSCL} distribution fits the data quite well compare to skew slash density. \textit{SSCL} captures skewness, peakedness and heavy tails in microarray data. Hence the \textit{SSCL} distribution provides the possibility of modelling impulsiveness and skewness required for gene expression data.

### 3.2.2 Application of Skew-slash Normal-Laplace Distribution

Here we illustrate the application of one of the distribution included in the family of Skew slash distributions generated by normal Kernal, Skew slash normal-Laplace (\textit{SSNL}) distribution to microarray data. We downloaded the cDNA dual dye microarray dataset (Experiment id-51398) from the Stanford Microarray Database. Each array chip contains approximately 42000 human cDNA elements, representing over 30000 unique genes. The dataset was normalized using Locally Weighted
CHAPTER 3. APPLICATION OF SKEW SLASH FAMILY OF DISTRIBUTIONS TO MICROARRAY DATA

Figure 3.1: MA plots of intensities from microarray Experiment id-38067 (a) LOWESS spline of intensities before loess normalization, (b) After LOWESS normalization.

Figure 3.2: Histogram of microarray gene expression data. The lines represent distributions fitted using maximum likelihood estimation: SSCL probability density function (red dashed line), skew slash (green line) and normal (black dotted line).
CHAPTER 3. APPLICATION OF SKEW SLASH FAMILY OF DISTRIBUTIONS TO MICROARRAY DATA

Figure 3.3: Box plots of intensities from microarray Experiment id-38067 (a) Before normalization, (b) After loess normalization.

Table 3.1: Microarray data analyzes - Maximum likelihood estimates and their asymptotical standard deviations for normal (N), skew slash normal (SSN), and skew slash Cauchy Laplace (SSCL) distributions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>SSN</th>
<th>SSCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>-0.002 (0.012)</td>
<td>0.072 (0.123)</td>
<td>0.062 (0.051)</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.832 (0.024)</td>
<td>0.598 (0.013)</td>
<td>1.881 (0.085)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-</td>
<td>-</td>
<td>0.618 (0.107)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-</td>
<td>0.362 (0.047)</td>
<td>0.490 (0.114)</td>
</tr>
<tr>
<td>$q$</td>
<td>-</td>
<td>5.27 (0.187)</td>
<td>8.15 (0.132)</td>
</tr>
</tbody>
</table>
Chapter 3. Application of Skew Slash Family of Distributions to Microarray Data

Figure 3.4: Histogram of microarray gene expression data. The lines represent distributions fitted using maximum likelihood estimation: *SSNL* probability density function (red dashed line), skew slash (green line) and normal (black dotted line).

Table 3.2: Microarray data analysis - Maximum likelihood estimates and their asymptotical standard deviations for normal (*N*), skew slash normal (*SSN*), and skew slash normal Laplace (*SSNL*) distributions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>SSN</th>
<th>SSNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.009 (0.004)</td>
<td>0.015 (0.003)</td>
<td>0.018 (0.021)</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.574 (0.004)</td>
<td>0.511 (0.013)</td>
<td>0.321 (0.078)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>-</td>
<td>-</td>
<td>0.561 (0.013)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>-</td>
<td>0.619 (0.047)</td>
<td>0.450 (0.094)</td>
</tr>
<tr>
<td>$q$</td>
<td>-</td>
<td>6.27 (0.187)</td>
<td>5.89 (0.082)</td>
</tr>
</tbody>
</table>

Linear Regression (LOWESS) (Cleveland and Delvin, 1988). We fitted the *SSNL* to the normalized microarray intensities. The estimated MLEs and standard errors (SE) are computed using the `optim` function of the R statistical software and is given in Table 3.2. Figure 3.4 depicts the histogram of the gene expression data and the fitted *SSNL* along with skew slash probability density function. We compared the empirical distribution of the microarray gene expression and the *SSNL* distribution. It can be clearly seen that the estimated density of the *SSNL* distribution fits the data quite well compare to skew slash density. *SSNL* captures skewness, peakedness and heavy tails in microarray data. Hence the *SSNL* distri-
distribution provides the possibility of modelling impulsiveness and skewness required for gene expression data compared to skew slash and normal.

### 3.2.3 Application of Skew Slash t Distribution

In this section, we present applications of the skew-slash $t$ ($SST$) distribution to cDNA dual dye microarray dataset (Experiment id-51398) from the Stanford Microarray Database. We estimated the parameters using maximum likelihood estimation using the `optim` function of the R statistical software, applying the BFGS algorithm (R Development Core Team (2006)). We fitted the skew-slash $t$ distribution to the normalized microarray intensities as given in Figure 3.5. The estimated MLEs and standard errors (SE) for the parameters are $\hat{\mu} = 0.018$ (0.003), $\hat{\sigma} = 0.421$ (0.038), $\hat{\alpha} = 1.605$ (0.106), $\hat{\nu} = 15.024$ (0.026) and $\hat{q} = 9.25$ (0.064).

We compared the empirical distribution of the microarray gene expression data with the skew-slash $t$ along with skew-slash and normal distributions evaluated at the MLEs. It can be clearly seen that the estimated density of the $SST$ fits the data quite well compared to skew-slash and normal densities. $SST$ captures skewness, peakedness and heavy tails.

### 3.2.4 Application of Asymmetric Slash Laplace

In this section we illustrate the application of asymmetric slash Laplace distribution to microarray data (Experiment id-51398) from the Stanford Microarray Database. Figure 3.6 depicts the histogram of the gene expression data and the fitted probability density function evaluated at the MLEs. We compared the empirical distribution function of the microarray gene expression data with the asymmetric slash Laplace ($ASL$), skew-slash and normal distributions evaluated at the MLEs.

It can be clearly seen that the estimated density of $ASL$ fits the data quite well compared to skew-slash and normal density. $ASL$ captures skewness, peakedness and heavy tails. Hence, provides the possibility of modelling impulsiveness and skewness required for gene expression data. From figure 3.6 it is observed that the gene expression data are asymmetric and the asymmetric slash Laplace dis-
Figure 3.5: Histogram of microarray gene expression data. The lines represent distributions fitted using maximum likelihood estimation: Skew slash t probability density function (red dashed line), skew slash (green line) and normal (black dotted line).

Figure 3.6: Fitted asymmetric slash Laplace (ASL) probability density function (red dashed line) to the microarray gene expression data along with skew slash (green line) and normal (black dotted line).
CHAPTER 3. APPLICATION OF SKEW SLASH FAMILY OF DISTRIBUTIONS TO MICROARRAY DATA

Figure 3.7: Fitted asymmetric slash Laplace (ASL) probability density function (red line) to the microarray gene expression data along with skew slash t (SST) (blue dashed line), skew slash Cauchy-Laplace (SSCL) (green dot line) and skew slash normal-Laplace (SSNL) (black dash dot line).

The fitted distribution describes the data well. The parameter estimates of the ASL together with the standard errors (given in brackets) are \( \hat{\theta} = 0.018 (0.003) \), \( \hat{\sigma} = 0.384 (0.016) \), \( \hat{\kappa} = 1.095 (0.023) \) and \( \hat{q} = 15.15 (0.010) \). It can be clearly seen that the estimated density of the ASL distribution fits the data quite well compared to skew slash and normal densities. It captures skewness, peakedness and heavy tails.

3.2.5 Comparison of Slash Family of Distributions

In this section we compared the performance of skew-slash Cauchy-Laplace (SSCL), skew slash normal-Laplace (SSNL), skew slash t (SST) and asymmetric slash Laplace (ASL) distribution. Figure 3.7 depicts the histogram of the gene expression data (Experiment id-51398) and the fitted probability density function evaluated at the MLEs. The estimated values for the parameters are given in Table 3.3. We compared the empirical distribution function of the microarray gene expression data with the ASL, SSCL, SSNL and SST distributions evaluated at the MLEs. It can be clearly seen that the estimated density of the ASL and SSCL fits the data quite well compared to skew slash normal-Laplace and skew-slash t. Both ASL and SSCL captures skewness, peakedness and heavy tails.
We have used Akaike’s Information Criterion (AIC) (Akaike (1973), Burnham and Anderson (1998)) to evaluate the comparative appropriateness of ASL. Let $f(\theta)$ is our model, then $AIC$ is given by

$$AIC = -2\log(L_f(\hat{\theta}|x_1,...,x_n)) + 2K,$$

where $K$ is the number of parameters being estimated, $L$ is the likelihood function of the model $f(\cdot)$, and $\hat{\theta}$ is the maximum likelihood estimate of the parameters of $f$. A smaller value of $AIC$ indicates a better fit. We calculated the $AIC$ values for the ASL, SSCL, SSNL and SST distributions. We found that $AIC_{ASL} - AIC_{SSCL} < 0$, $AIC_{ASL} - AIC_{SSNL} < 0$ and $AIC_{ASL} - AIC_{SST} < 0$, which implies the better fit for the ASL. Hence the asymmetric slash Laplace distribution with smaller value of $AIC$ is the better model for microarray gene expression data.

We estimated maximum likelihood estimates and asymptotic standard errors of the parameters of a ASL distribution for the arrays. The estimated value of $\kappa$ is 1.095, which is close to one across the cDNA arrays, indicating small levels of skewness. Estimate of $\theta$ equal to 0.018, gives the measure of the center of the gene expression values and the estimate of $\sigma$ equal to 0.384, gives the scale. Hence ASL distribution can give parametric insight into normalization across arrays. Use of MLE estimates of $\theta, \sigma$ and $\kappa$, allow for easier comparison among the
arrays because these estimates can account for the different skewness of different arrays in evaluating proper measures of center and scale. Hence the ASL model nicely separates the location parameter from the skew parameter, so the effect of the two can be taken into account in determining what further normalization analysis is appropriate.

3.3 Interpretation of ASL as Error Distribution

From the Figure 3.6 we can see that ASL accounts the skewness, peakedness and heavier tails in the microarray data. From the Eq. (2.5.8) we can see that ASL can be represented as a continuous mixture of normal random variables whose scale and mean parameters are dependent and vary according to an exponential distribution. Giles and Kipling (2003) examined the distribution of genes across arrays for oligonucleotide microarrays using 59 replicated arrays of the same sample. They found the distribution of a gene's expression follows a normal distribution, with random standard deviation and mean from an exponential distribution. Due to the nature of microarray experiments, we would expect the measured intensities to have different variation across genes, and a mixture of normals is a convenient representation. This justifies the use of ASL as the error model for microarray.

Also arrays are often measured as log-ratios of the red and green channel, then the scale mixture of log-ratio of Pareto representation given in Eq. (2.5.9) gives the explanation of the good fit of the ASL to the data, if the red and green channel each follow independent Pareto distributions. Kuznetsov (2001) found mRNA expression in SAGE libraries following a Pareto-like distribution. Similarly, Wu et al. (2003) found that the distribution of the expression intensities for Affymetrix oligonucleotide arrays resemble a power law, which is equivalent to a Pareto distribution. The ASL distribution has power tails like Pareto distribution. These characteristics of ASL makes this distribution suitable for modelling microarrays.
3.4 Conclusion

In this chapter we illustrated the applications of new families of skew slash distributions introduced in chapter 2. These new families of skew slash distributions have wide flexibility in tail behavior, skewness and peakedness and can be applied to data from different contexts especially for heavy-tailed and skewed data sets. 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 skew slash distributions. The skew slash distributions captures the peakedness at the center of the data, heavier tails as well as the asymmetry in the distribution. These distributions gives parametric insight into normalization across arrays and can be used as an error model.

Heavy-tailed distributions are commonly found in complex multi-component systems like ecological systems, biometry, economics, engineering, insurance, genetics, meteorology, hydrology, climatology, sociology, internet traffic, file size stored on a server, financial return data, telecommunications, signal processing, environmental data, reliability and risk data, etc. Slash distributions are potentially useful for empirical modeling in these situations since it provides the flexibility for modelling impulsiveness and skewness observed in these datasets. One of the disadvantage for these distributions is that the estimation of the parameters are computationally complicated. This motivated us to introduce another distribution called asymmetric type II Laplace ($ACL$) distribution which is computationally less complicated compared to the skewed slash family. The $ACL$ have asymmetry, peakedness and heavier tails like the skew slash family.

References


Bindu, P. P. (2011b). A new family of skewed slash distributions generated by the


82
CHAPTER 3. APPLICATION OF SKEW SLASH FAMILY OF DISTRIBUTIONS TO MICROARRAY DATA
