Application of skew-normal distribution for detecting differential expression to microRNA data

Ahmed Hossain\textsuperscript{abc} & Joseph Beyene\textsuperscript{ab}

\textsuperscript{a} Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada;
\textsuperscript{b} Statistics for Integrative Genomics and Meta-Analysis, McMaster University, Hamilton, Canada;
\textsuperscript{c} Pharma Medica Research Inc., Mississauga, ON, Canada

Published online: 29 Sep 2014.
Application of skew-normal distribution for detecting differential expression to microRNA data

Ahmed Hossain\textsuperscript{a,b,c}\ and Joseph Beyene\textsuperscript{a,b}\textsuperscript{*}

\textsuperscript{a}Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada; \textsuperscript{b}Statistics for Integrative Genomics and Meta-Analysis, McMaster University, Hamilton, Canada; \textsuperscript{c}Pharma Medica Research Inc., Mississauga, ON, Canada

(Received 3 October 2013; accepted 3 September 2014)

Traditional statistical modeling of continuous outcome variables relies heavily on the assumption of a normal distribution. However, in some applications, such as analysis of microRNA (miRNA) data, normality may not hold. Skewed distributions play an important role in such studies and might lead to robust results in the presence of extreme outliers. We apply a skew-normal (SN) distribution, which is indexed by three parameters (location, scale and shape), in the context of miRNA studies. We developed a test statistic for comparing means of two conditions replacing the normal assumption with SN distribution. We compared the performance of the statistic with other Wald-type statistics through simulations. Two real miRNA datasets are analyzed to illustrate the methods. Our simulation findings showed that the use of a SN distribution can result in improved identification of differentially expressed miRNAs, especially with markedly skewed data and when the two groups have different variances. It also appeared that the statistic with SN assumption performs comparably with other Wald-type statistics irrespective of the sample size or distribution. Moreover, the real dataset analyses suggest that the statistic with SN assumption can be used effectively for identification of important miRNAs. Overall, the statistic with SN distribution is useful when data are asymmetric and when the samples have different variances for the two groups.

\textbf{Keywords:} MicroRNA expression data; skew-normal distribution; false discovery rate; \texttt{sn} R package

\textbf{AMS Classification:} Statistics

\section{Introduction}

The skew-normal (SN) distribution offers a more flexible formulation compared with the normal distributions by bringing a shape parameter, which regulates skewness. This distribution was introduced and popularized by Azzalini [4,5] and has been discussed in detail with variations by other authors [14,19]. Recently, Azzalini [6] presented a discussion on SN distributions with

\textsuperscript{*}Corresponding authors. Emails: ahmed.hossain@utoronto.ca; beyene@mcmaster.ca

© 2014 Taylor & Francis
applications in regression models. Extensions, further properties and applications of the SN distribution can also be found in Genton [18]. During the last few years, a substantial development of the theory of SN distributions is observed and applied to various fields [14,19,28,33,45]. Here, our aim is to describe the scope of applicability of the SN distribution to analyze microRNA (miRNA) expression data.

miRNAs are short non-coding RNAs that play critical roles in numerous cellular processes affecting gene expression [2,3,21]. The aberrant role of miRNAs has been reported in a number of diseases such as cancer [34] and heart disease [44]. A single miRNA has multiple target genes and, thus, could control a large number of protein-coding genes. This may explain why miRNAs play an important role in the regulation of diverse cellular processes. Increasing efforts to identify specific targets of miRNAs have led to the speculation that miRNAs may regulate at least 30% of human protein encoding genes. Therefore, it is essential to apply a robust computational and statistical methodology to identify differentially expressed miRNAs between two or more groups. In such analyses, the main purpose is to determine whether or not the list of miRNAs is active in a certain disease. There are often hundreds of miRNAs tested at a time, but there might be few miRNAs that are actually active for a certain disease [27]. If a statistical method is reasonably good, most of the active miRNAs would appear in the list of top ranked miRNAs.

The linear regression models, based on normal distribution assumption, play an essential role in miRNA expression analysis. However, there is an indication that the normality assumption does not work well in certain situations, being specially sensitive to the presence of noise in the data which is very common in miRNA expression. In fact, several authors showed that gene expression data from microarrays are often not normally distributed, even after some preprocessing [12,26,30,41,46,47]. More recently, Hossain et al. [23] used a generalized logistic distribution of Type II and Bhowmick et al. [11] used a Laplace mixture model approach for identification of differentially expressed genes in microarray experiments. Lonnstedt and Speed [29] proposed a normal mixture model for gene expression data and defined a log posterior odds statistic. Alternatively, distribution-free methods that are not based on any parametric assumption can be used [22,24,25,31]. It is now a common practice to control multiple testing when identifying differentially expressed miRNAs. We used the smother method described in Storey and Tibshirani [42] to estimate the false discovery rate (FDR) and these FDR values are used to rank the miRNAs.

The present paper contains three main contributions: (1) we apply SN distribution for detecting a list of target miRNAs. This assumption provides a long-tailed alternative to the normal distribution, (2) we propose a Wald test statistic for analyzing miRNA data that can also be used in analyzing microarray gene expression data, RNA sequence data, protein expression data and genome-wide association studies with quantitative traits, etc., (3) a comparison of the methods that provides their strengths and limitations at skewed settings of simulation.

2. Motivation

It is usual practice in the analysis of miRNA expression data to assume that the underlying distribution for expression levels is normal, although this assumption may not always support the analysis. This motivates us to use the SN distribution which has longer tails than normal distribution. Figure 1 displays distributions (Gaussian kernel smoothing by the density() function in R) of four randomly selected miRNAs from the high-risk Neuroblastoma miRNA expression dataset. Details about the dataset are given in Scaruffi et al. [38]. Our interest here is to build a separation between two survivor types: short (death within 36 months from diagnosis) and long (> 36 months). It appears from the figure that the underlying distributions of miRNAs lack symmetry and the degree of skewness for each miRNA varies. We examine the likelihood-ratio (LR) test for testing if the shape parameter of the SN distribution is equal to 0 for the long survivor group of each miRNA, by reference to a chi-square distribution with 1 degree of
After investigation of the $p$-values, we have seen that most (72%) of the $p$-values are less than 0.05 which confirms that the SN distribution might be promising for research in miRNA study. Therefore, in the context of miRNA studies it is observed that many miRNAs have skewed distributions, which is why in those miRNAs the normal distribution is an inadequate assumption for analyzing the data. This motivates us to apply the SN distribution, which may be a better fit for the expression profiles than a normal distribution.

3. Skew-normal distribution

3.1 Distribution

A random variable $Y$ has a SN distribution with location parameter $\mu$, scale parameter $\sigma$ and shape parameter $\lambda$, if the pdf of $Y$ is given by

$$f_Y(y \mid \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi \left( \frac{y - \mu}{\sigma} \right) \Phi \left( \lambda \frac{y - \mu}{\sigma} \right),$$

where $\phi$ and $\Phi$ are the probability density function (pdf) and the cumulative distribution function (cdf), respectively, of the $N(0, 1)$. We describe this by using the notation $Y \sim SN(\mu, \sigma^2, \lambda)$. For $\lambda = 0$ the pdf corresponds to the normal distribution with mean $\mu$ and variance $\sigma^2$. The mean
and variance for SN distribution are given, respectively, by

\[ E(Y) = \mu + c\sigma \frac{\lambda}{\sqrt{1 + \lambda^2}}, \]
\[ \text{Var}(Y) = \sigma^2 \left(1 - c^2 \frac{\lambda^2}{1 + \lambda^2}\right), \]

where \( c = \sqrt{2/\pi} \). Sahu et al. [35] defined the SN distribution with skewness parameter \( \delta \) as

\[ f_Y(y | \mu, \sigma^2, \delta) = \frac{2}{\sqrt{\sigma^2 + \delta^2}} \phi \left( \frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} \right) \Phi \left( \frac{\delta}{\sigma} \frac{y - \mu}{\sqrt{\sigma^2 + \delta^2}} \right). \]

Sahu et al. [35] obtained the shape parameter \( \lambda \) from the skewness parameter, that is, \( \delta = \lambda/(\sqrt{1 + \lambda^2}) \). The shape parameter can model right and left skewness, and when the shape parameter is zero, the model reduces to the normal distribution. Moreover, it is mathematically tractable and its moment generating function has a closed form.

3.2 Estimation of parameters

Azzalini [4] provided the Fisher information matrix for the direct parameterization of a SN distribution. Recently, Hallin and Ley [20] provided the score and Fisher information matrix for the skew-symmetric probability density function. Analytical expressions for the maximum likelihood estimates (MLEs) of location and scale parameters can be obtained, for a given scale parameter, and the scale parameter can then be estimated by the profile likelihood method. Azzalini and Capitanio [9] gave the Fisher information matrix for the mean, variance and skewness parameter for the SN distribution. They used an application of the delta method for estimating the standard errors (SEs) for these three parameters.

The ‘sn.mle()’ function from the R package sn provides the MLE of the SN parameters [7]. But sometimes estimation of the parameters of the SN distribution using this MLE procedure encounters difficulties. Very often, the shape parameter of the SN distribution diverges especially with small sample sizes. To avoid this situation, an alternative estimation criterion is the method of Sartori-Firth [16,36], which involves first regular MLE and subsequent re-estimation of the shape parameter using a modified score function. Since each evaluation of the correction term involves two numerical integrations, computations are carried out by adopting an approximation due to Bayes and Branco [10]. The ‘sn.mmle()’ function from the R package sn provides the parameter estimation using a modification of MLE. The theory of ‘sn.mmle()’ is given in Azzalini [8] where he used a penalized likelihood estimation method to avoid the difficulties of convergence during the estimation. In our miRNA analysis we used the sn.mle() function from the sn package but if it did not successfully converge we used sn.mmle() function for parameter estimation. If neither of the functions is successful in estimating the parameters, then we can use the normal distribution assuming that the shape parameter is 1. Another limitation of using the sn package is that it does not handle missing values. Therefore, it is necessary to have a complete dataset before applying the function.

Figure 2 presents the histograms of estimates for the SN parameters using the Neuroblastoma dataset. We used the ‘sn.mle()’ function from the R package sn for parameter estimation. We found that 72 (22.5%) and 66 (20.6%) miRNAs have unsuccessful completion of convergence for long and short groups, respectively. This problem has received some attention for the (limiting case of the) SN distribution for boundary estimates of the shape parameter [32]. In such situation, we used the ‘sn.mmle()’ function. It appears from the histograms that the location parameter for most miRNAs lies close to 2 and the scale parameter lies between 0 and 6. It is
also seen that the scale parameter for the short survivor group has many high values as indicated by comparison with upper quartile values. The histogram of the shape parameters (results from sn.mmle() function) indicates that the estimates of the shape parameters lie between $-20$ and $20$. Therefore, some miRNAs exhibit heavy tails. The SN distribution is therefore found promising since they preserve the advantages of the normal distribution with the additional benefit of flexibility by considering skewness and kurtosis.

### 3.3 Wald test for comparing means of two groups

Let us define the mean of the treatment group as $M_1$ and the variance of the mean as $\text{var}(M_1)$. Similarly, we define the mean of the control group as $M_2$ and the variance of the mean as $\text{var}(M_2)$. Now for testing the equality of the means between two groups, the test statistic, $Z_g$, for $g$th miRNA becomes

$$Z_g = \frac{M_1 - M_2}{\text{SE}(M_1 - M_2)}$$

where

$$\text{SE}(M_1 - M_2) = \sqrt{\text{var}(M_1) + \text{var}(M_2)},$$
which is approximately standard normally distributed. We can rank miRNA according to the values of $Z_g$.

However, when there are only a small number of arrays in each group, the estimates of SE for each miRNA can be unstable. Some miRNAs might by chance have very small SEs, and therefore appear highly significant. To address this problem, we smooth the variance estimates by borrowing information from the ensemble of miRNAs. This can assist in inference about each miRNA individually. This technique of smoothing variances is not new in microarray studies. For example, [15,43] and [13] used $t$-statistics where an offset was added to the standard deviation while [39] proposed a $t$-statistic with a Bayesian adjustment to the denominator. We took the offset $s_0$ as the 90th percentile of the miRNA-wise SEs following Efron et al. [15]. Therefore, we can calculate the $D_g$ statistic to test for treatment effect as

$$D_g = \frac{M_1 - M_2}{SE(M_1 - M_2) + s_0}.$$

Similar adjustments for computing a test statistic were also used by Garrett-Mayer et al. [17] and Hossain et al. [25]. The R code for testing the difference between two group means is provided in the appendix.

### 3.4 False discovery rate estimation with SN-D statistic

In miRNA analysis, multiplicity arises due to testing thousands of hypotheses. FDR of a statistical method is commonly used in multiple hypotheses testing problem to correct for multiple comparisons. The procedures are designed to control the expected proportion of false positives among the declared significant results. FDR is estimated using permutation and giving a threshold to the statistic [24,42]. We can use the following permutation algorithm to select differentially expressed miRNAs and estimate FDR:

1. For the $g$th miRNA, calculate the SN-$D_g$ statistics, denote their ordered values as $D_{(g)}$.

2. Permute the samples of the miRNA expression data and recalculate the $D_g$ statistic. For the $b$th permutation, calculate the $D_g^b$ statistic and denote their ordered values as $D_{(g)}^b; b = 1, \ldots, B$.

   Denote their averages across all permutations as

   $$\bar{D}_{(g)} = \frac{1}{B} \sum_{b=1}^{B} D_{(g)}^b.$$  

3. For a threshold, $\Delta$, identify the following miRNAs as significant:

   $$|D_{(g)} - \bar{D}_{(g)}| \geq \Delta.$$

   Denote $D_0 = \max_{D_{(g)} \leq \bar{D}_{(g)} - \Delta} D_{(g)}$ and $D_1 = \min_{D_{(g)} \geq \bar{D}_{(g)} + \Delta} D_{(g)},$ and estimate the expected number of false positives by chance for the $D_g$ statistic as follows:

   $$V(\Delta) = \frac{1}{B} \sum_{b=1}^{B} \left( I[D_{(g)}^b \geq D_1] + I[D_{(g)}^b \leq D_0] \right).$$
where \( I(\cdot) \) is the indicator function, and the estimated FDR is

\[
\hat{\text{FDR}}(\Delta) = \frac{V(\Delta)}{R(\Delta)},
\]

where

\[
R(\Delta) = \sum_g I(|D(g) - \bar{D}(g)| \geq \Delta)
\]

is the total number of significant genes. We can similarly calculate the expected number of false positives and FDR for the \( t \) and rank sum test (RST) statistics.

### 3.5 Comparison with other methods

We evaluate the performance of the SN-D statistic using the simulated data as well as two published datasets. Here, we consider \( t \)-statistic, RST and empirical Bayes \( t \)-statistics (moderated-\( t \), Mod-\( t \)) for comparison because these methods are widely used methods in the miRNA selection literature. The Mod-\( t \) statistic is implemented in the R/LIMMA package.

### 3.6 Simulation results

The performance of the methods is evaluated using the simulated data generated from two distributions incorporating variability, treatment effect and sample size effect. We consider the scenarios having two conditions, treatment versus control, and sample sizes of 20 and 40 per condition. We generated data from 1000 miRNAs and considered the proportion of differentially expressed miRNAs as 0.1. The following simulation scenarios are considered here:

- **Sim1**: Sim1 generates data from the SN distribution with different scale and shape parameters.
  - **Location**: We set the location parameter as 2.
  - **Scale**: We consider the scale parameters as 1.25, 1.5, 2 and 2.5 for the treatment group and 1.25 for the control group. This allows us to investigate the performance of the methods for both equal and unequal variances of the two groups.
  - **Shape**: We consider the shape parameters as 0, 1, 2, 5 and 8 for each group. The shape parameter 0 allows us to see the performance of the methods at normal distribution. We consider the treatment effect size as 0.5 which is added to the treatment group for the first 10% of the miRNAs.

- **Sim2**: Sim2 generates data from the extreme value (EV) distribution. This distribution allows us to see the performance of the methods in the presence of the smallest and largest expression values in the data. We consider the location and scale parameters as follows:
  - **Location**: We set the location parameter as 0.
  - **Scale**: We consider scale parameters as \( b' = 1, 2, 3 \) and 4 for treatment group and \( b' = 1 \) for control group. It allows EVs at the right side of the distribution with different variances for two groups. That is expressions for a miRNA are generated from \( \text{EV}_m(0, b') \). Therefore, the mean (i.e. \( b'\gamma \), where \( \gamma \) is Euler’s constant) and variance (i.e. \( b'^2(\pi^2/6) \)) for each miRNA within a group depends only on scale parameter. We consider the treatment effect size as 1 which is added to the treatment group for the first 10% of the miRNAs.
A. Hossain and J. Beyene

Figure 3. Average number of true positive miRNAs based on 1000 simulations corresponding to different shapes of the SN distribution and sample sizes are 20 per group. (a) Treatment groups are simulated from SN distribution with fixed scale parameter as 1.25, (b) treatment groups are simulated from SN distribution with fixed scale parameter as 1.5, (c) treatment groups are simulated from SN distribution with fixed scale parameter as 2.0 and (d) treatment groups are simulated from SN distribution with fixed scale parameter as 2.5.

The number of true positives is estimated by taking the average number of miRNAs that are correctly identified from the set of 100 top ranked miRNAs based on 1000 simulations. We consider ranking miRNAs by the FDR values. Figures 3 and 4 show the average number of miRNAs that are correctly identified from the set of 100 top ranked miRNAs by each of the methods. For example, when applying the SN-D statistic to the simulated dataset with shape parameter 5, scale parameter 1.5 and sample size 20 per group, we observe that on average, 54.68 miRNAs are correctly identified from the list of top 100 ranked miRNAs. Moreover, the t-statistic, RST and Mod-t statistic produced on average, 52.94, 52.12 and 54.02 miRNAs, respectively. On the other hand, the numbers are 73.44, 75.44, 74.80 and 76.68 for RST, Mod-t, t statistics and SN-D statistics, respectively, at sample size 40 per condition. It appears from the result that the D statistic with the assumption of SN distribution performs best among all methods when the sample size is considered large, unequal variances among groups and presence of skewness in the data. Again, when the expression of a miRNA comes from the normal distribution (i.e. shape parameter at 0), the SN-D statistic and the Mod-t statistic provide similar results.
Figure 4. Average number of true positive miRNAs based on 1000 simulations corresponding to different shapes of the SN distribution and sample sizes are 40 per group. (a) Treatment groups are simulated from SN distribution with fixed scale parameter as 1.25, (b) treatment groups are simulated from SN distribution with fixed scale parameter as 1.5, (c) treatment groups are simulated from SN distribution with fixed scale parameter as 2.0 and (d) treatment groups are simulated from SN distribution with fixed scale parameter as 2.5.

Table 1. Average number of miRNAs truly differentially expressed from the top ranked 100 miRNAs after 1000 simulations of the EV distribution.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Sample size</th>
<th>SD</th>
<th>RST</th>
<th>Mod-$t$</th>
<th>$t$-stat</th>
<th>SN-$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV(0,1)</td>
<td>$n_1 = n_2 = 20$</td>
<td>1.28</td>
<td>64.87</td>
<td>66.18</td>
<td>65.69</td>
<td>65.32</td>
</tr>
<tr>
<td>EV(0,2)</td>
<td>$n_1 = n_2 = 20$</td>
<td>2.56</td>
<td>50.72</td>
<td>51.7</td>
<td>51.8</td>
<td>51.28</td>
</tr>
<tr>
<td>EV(0,3)</td>
<td>$n_1 = n_2 = 20$</td>
<td>3.85</td>
<td>37.47</td>
<td>38.74</td>
<td>38.82</td>
<td>39.11</td>
</tr>
<tr>
<td>EV(0,4)</td>
<td>$n_1 = n_2 = 20$</td>
<td>5.13</td>
<td>30.42</td>
<td>30.96</td>
<td>31.86</td>
<td>31.91</td>
</tr>
<tr>
<td>EV(0,1)</td>
<td>$n_1 = n_2 = 40$</td>
<td>1.28</td>
<td>84.02</td>
<td>85.48</td>
<td>84.04</td>
<td>84.43</td>
</tr>
<tr>
<td>EV(0,2)</td>
<td>$n_1 = n_2 = 40$</td>
<td>2.56</td>
<td>51.47</td>
<td>52.02</td>
<td>52.34</td>
<td>52.56</td>
</tr>
<tr>
<td>EV(0,3)</td>
<td>$n_1 = n_2 = 40$</td>
<td>3.85</td>
<td>39.52</td>
<td>40.04</td>
<td>40.46</td>
<td>40.57</td>
</tr>
<tr>
<td>EV(0,4)</td>
<td>$n_1 = n_2 = 40$</td>
<td>5.13</td>
<td>31.53</td>
<td>32.01</td>
<td>32.38</td>
<td>32.66</td>
</tr>
</tbody>
</table>

Note: SD, standard deviation for the treatment group; RST, rank sum test; Mod-$t$, empirical Bayes $t$ from LIMMA; SN-$D$, skew-normal $D$ statistic.
It appears from the figures that all the methods outperform the RST statistic in terms of selecting the number of true positive miRNAs. It is also observed that the performance of the SN-D statistic increases with the presence of skewness in the data. Comparing Figures 3 and 4, we noticed that the SN-D statistic performs best with large sample sizes for both groups. That is, focusing on differences between the two conditions with large samples yielded SN-D statistic a better selection algorithm especially with skewed data. Table 1 presents the results of the average number of truly differentially expressed miRNAs from the top ranked 100 miRNAs after 1000 simulations from the EV distribution. It appears from the top ranked results that the Mod-t statistic may be more efficient than the SN-D statistic when the two groups have homogeneous variances. But the gain in efficiency reduces when the variances for the two groups are different. It is also seen from the results that the SN-D statistic performs competitively with the t statistic or Mod-t statistic irrespective of sample size or distribution. It also appears from the results that the relative performance of SN-D statistic improves when the two groups have different variances. It should be noted that the presence of noise is very common in real miRNA data. Therefore, even when the normality assumption holds for a given miRNA data the SN-D statistic performs well for identifying differentially expressed miRNAs.

3.7 Scaruffi et al. Neuroblastoma data

In the present study, we evaluated the miRNA expression profile of 31 high risk, stage 4 Neuroblastoma patients. The expression data for 319 miRNAs is collected by using Agilent Technologies [1]. We compared miRNA expression profiles of 14 long-survivors (alive with an overall survival time > 36 months) and 17 short-survivors (dead of disease within 36 months) from diagnosis. Data were collected from GEO and they are accessible through GEO Series accession number GSE16444. Further details about the dataset is given in the paper by Scaruffi et al. [38]. The results of 250 miRNAs which were ranked according to the Mod-t statistic are presented in the GEO website. We compare the methods in terms of concordance of miRNAs from the list of top ranked 50 miRNAs. Concordance is defined as the number of miRNAs in one selected list produced by one method, which are also present in another miRNA list produced by another method. Figure 5(a) gives a Venn diagram of the top 50 ranked miRNAs by t-statistic, Mod-t statistic (LIMMA), RST and SN-D statistic. It appears that the SN-D statistic provides 38 miRNAs which are commonly found by Mod-t (LIMMA) statistic.

![Figure 5](image.png)

Figure 5. Comparison of four statistical methods with Neuroblastoma data. (a) Venn diagram of top 50 ranked miRNAs by four methods and (b) ROC plot by top 20 miRNAs predictors.
Table 2. Average of classification errors (with their SEs in parentheses) for colon cancer dataset.

<table>
<thead>
<tr>
<th>$k_0$</th>
<th>Mod-$t$</th>
<th>$t$-stat</th>
<th>RST</th>
<th>SN-$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.241 (0.134)</td>
<td>0.259 (0.142)</td>
<td>0.265 (0.159)</td>
<td>0.248 (0.141)</td>
</tr>
<tr>
<td>10</td>
<td>0.213 (0.128)</td>
<td>0.234 (0.165)</td>
<td>0.237 (0.156)</td>
<td>0.216 (0.131)</td>
</tr>
<tr>
<td>15</td>
<td>0.199 (0.126)</td>
<td>0.226 (0.158)</td>
<td>0.220 (0.158)</td>
<td>0.211 (0.129)</td>
</tr>
<tr>
<td>20</td>
<td>0.158 (0.132)</td>
<td>0.178 (0.171)</td>
<td>0.179 (0.145)</td>
<td>0.163 (0.136)</td>
</tr>
<tr>
<td>25</td>
<td>0.127 (0.125)</td>
<td>0.156 (0.157)</td>
<td>0.145 (0.146)</td>
<td>0.129 (0.126)</td>
</tr>
<tr>
<td>30</td>
<td>0.115 (0.140)</td>
<td>0.148 (0.162)</td>
<td>0.132 (0.148)</td>
<td>0.119 (0.141)</td>
</tr>
</tbody>
</table>

Note: $k_0$ is the number of top ranked miRNAs.

In addition, we fitted logistic models to the list of top ranked 20 miRNAs that are found by the four methods and estimated the prediction of a probability of a ‘success’ (here, long survivors being in case group). For example, the list of top ranked 20 miRNAs by $t$-statistic were considered as explanatory variables to predict the short- and long-survivor group by fitting a logistic regression model. These predicting binomial data are evaluated by receiver operating characteristic (ROC) curves, which are shown in Figure 5(b). It appears from the figure that the SN-$D$ statistic provides better discrimination of the two conditions than any of the other methods. Therefore, in this particular example, the SN-$D$ statistic identified differentially expressed miRNAs that may have more biological relevance from the list of top 20 ranked miRNAs.

3.8 Sarver et al. colon cancer data

Sarver et al. [37] analyzed the miRNA expression data of colonic tumor and healthy tissue obtained during surgery from 145 subjects [116 with colon tumor and 29 normal]. These data have also been analyzed by Hossain and Beyene [22]. Data were collected from GEO with accession number GSE18392. The main objective is to determine statistically significant miRNAs that are differentially expressed between normal tissue and all tumor tissues. The results of 250 miRNAs were ranked according to the Mod-$t$ statistic in the GEO website. Importantly, Sarver et al. [37] mentioned that miR-96, miR-182, miR-182* and miR-183 were all up regulated; expression was highly correlated and mapped to the same region of chromosome 7. Again, miR-135b displayed the largest average change, a 4.55-fold increase. It is noticed that the Mod-$t$ statistic is not able to identify the miRNA miR-96 from the list of first 20 miRNAs but our method can identify this miRNA if we choose to select the top ranked 20 miRNAs.

Here, we evaluate the effectiveness of the miRNA list by different methods to form a classifier which could predict the class of a test sample. In using a classification to obtain the best method, we assumed that a better miRNA list should discriminate between the groups more effectively. Therefore, we evaluate the classification performance of the four methods $t$-statistic, Mod-$t$, RST, and SN-$D$ statistic using a simple Gaussian maximum likelihood discriminant rule [40]. The form of the algorithm is as follows:

(1) Split the data into five folds (subsamples).
(2) Of the five folds, a single fold is retained as the validation data for testing the model, and the remaining four folds are used as the training data. In choosing the folds, we ensured that at least five subjects are chosen from each group.
(3) The cross-validation process is repeated for each of the folds, with each of the five folds used exactly once as the validation data.
(4) Record the number of top ranked genes from the training data and use these genes for classification with a simple Gaussian maximum likelihood discriminant rule. The error for a given classification relative to a known truth is calculated by the \texttt{classError} function of \textit{R} package \texttt{mclust}.

(5) Report the average error over all the test results.

The top ranking genes of a specified number ($k_0$) between 5 and 30 are used to create the classification rule. The mean of the misclassification errors and their corresponding SEs for different methods are summarized in Table 2. It appears that the SN-$D$ statistic performs similar to Mod-$t$ statistic choosing minimum classification errors. In fact, compared to other methods, the Mod-$t$ and SN-$D$ statistics produce more consistent results. Therefore, the Mod-$t$ and SN-$D$ statistics can be used effectively to this dataset for identification of the important miRNAs.

4. Discussion and conclusion

The aim of this paper was to investigate the flexibility of the SN distribution to identify differentially expressed miRNAs. It is of special interest from the applied viewpoint to have at hand the SN distribution for which we can regulate both skewness and thickness of the tails. Among the various alternatives, an appealing option is offered by a skewed version of the normal density, since this is fairly tractable from the algebraic viewpoint and its symmetric version. An important feature of this distribution is that it accommodates both the skewness and the normal density.

This paper is motivated by the belief that SN distribution might lead to robust results in miRNA data since extreme outliers are very common in such data. We bring the techniques to use SN distribution relaxing the normality assumption for miRNA data. For our analysis, we adopt a Wald-type framework to develop a SN-$D$ statistic for identifying differentially expressed miRNAs. The simulation results suggested that the performance of SN-$D$ statistic improves with larger sample sizes and performed best when the data exhibits skewness and the two groups had different variances. It also appeared that the SN-$D$ statistic performs comparably with other statistics irrespective of sample size or distribution. We also demonstrated our approach with two real datasets and it appears that the statistic with SN assumption performed competitively with other methods.

The use of SN distribution in miRNA data analysis has limitations. There are more parameters needed to estimate in applying SN-$D$ statistic than normal assumption. A solution of the estimation problem is recommended in the ‘sn /R’ package though there are some issues involved in estimating the shape parameter. The difficulties associated with parameter estimation with SN distribution often arises with small sample sizes. Therefore, it is not recommended to use this distribution when sample sizes are very small (e.g. $n = 10$ or less per group). Nonetheless, the SN assumption is attractive in miRNA applications, because it can fit both symmetric and asymmetric data according to expression values. The shape or skewness parameter highlights the directions for a miRNA where departures from normality are more pronounced.

In summary, the $D$ statistic with SN assumption is a promising candidate to identify differentially expressed miRNAs. The presence of outliers or noise is very common in miRNA expression data and therefore the nature of the skewness that might arise in the miRNA data should be considered carefully. The proposed method contributes an alternative analytical approach to the analysis of such data. The method with SN assumption can also be applied to other recently developed high-throughput technologies, for example, analysis of RNA-sequence data, genome-scale inference problems such as genome-wide association studies for quantitative trait loci, etc.
Acknowledgements

A.H. acknowledges Canadian Institute of Health Research (CIHR) fellowship funding from the Drug Safety and Effectiveness Cross-Disciplinary Training (DSECT) Program. J.B. would like to acknowledge Discovery Grant funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) (grant number 293295-2009) and CIHR (grant number 84392). J.B. holds the John D. Cameron Endowed Chair in the Genetic Determinants of Chronic Diseases, Department of Clinical Epidemiology and Biostatistics, McMaster University. We would like to thank Professor Adelchi Azzalini for contributing the ideas of estimating parameters for high dimensional settings. We would also like to thank two anonymous reviewers and the editor for insightful comments that improved the presentation and clarity of our manuscript.

References

Appendix

```r
### The following R code provides the test statistic to compare the means of two skew normal distributions.
require("sn")
### The SN.Z.mmle() function estimates the mean and standard error of the mean for SN
### distribution using sn.mmle() function.
SN.Z.mmle<-function(y){
  # y vector of a miRNA expression for a group
```


n <- length(y)
b2 <- NULL
try.out<-try(sn.mmle(y=y, plot.it=FALSE),silent=TRUE)
if (class(try.out) == "try-error") {
  b2$dp <-list( Mean = NA, se = NA)
} else { b2 <- try.out
dp<-b2$dp
X <- as.matrix(rep(1, n))
colnames(X) <- "constant"
info<-sn.Einfo(dp = dp, x = X)
list( Mean = as.numeric(info$cp[1]), se = as.numeric(info$se.cp[1]))
}

### The sn.wald() return a p-value from the the Wald test statistic of testing equality
### of two means.
sn.wald<-function(g,L) {
  # g vector of a miRNA expression for two groups
  # L two groups defined as either 0 or 1
  g1<-as.numeric(g[L==0])
g2<-as.numeric(g[L==1])
g11<-g1[!is.na(g1)] # Delete the missing values
g22<-g2[!is.na(g2)] # Delete the missing values
bb1<-sn.mle(y=g11, plot.it=FALSE)
bb2<-sn.mle(y=g22, plot.it=FALSE)
meanx<-as.numeric(bb1$cp[1])
meany<-as.numeric(bb2$cp[1])
sdx<-as.numeric(bb1$se[1])
sdy<-as.numeric(bb2$se[1])

  mean1<-ifelse(bb1$optim$convergence==0,meanx,as.numeric(SN.Z.mmle(g11)[1]))
  mean2<-ifelse(bb1$optim$convergence==0,meany,as.numeric(SN.Z.mmle(g22)[1]))
  sel<-ifelse(bb1$optim$convergence==0,sdx,as.numeric(SN.Z.mmle(g11)[2]))
  se2<-ifelse(bb1$optim$convergence==0,sdy,as.numeric(SN.Z.mmle(g22)[2]))

  sdl.2<-sqrt(sel^2+se2^2)
  mean1.2<-mean1-mean2
  waldt<-mean1.2/sdl.2
  p.val<-2*pnorm(-abs(waldt))
  return(p.val)
}

### Example
> x1<-rsn(n = 100, location = 1, scale = 2, shape = 3)
> x2<-rsn(n = 100, location = 2, scale = 2, shape = 5)
> g<-c(x1,x2)
> L<-c(rep(1,length(x1)),rep(0,length(x2)))
> sn.wald(g,L)
[1] 5.835806e-08