Far East J. Theo. Stat. **18**(2) (2006), 145-160

STATISTICAL PROPERTIES AND A LIKELIHOOD RATIO TEST FOR A DISTORTED NORMAL MODEL

JIE CHEN*, JOHN T. CHEN and ARJUN K. GUPTA

*Department of Mathematics and Statistics
University of Missouri-Kansas City
Kansas City, MO 64110, U. S. A.

Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, Ohio 43403, U. S. A.

Abstract

This paper discusses a stochastic representation, the moment generating function, and a likelihood ratio test for a distribution family that extends the skew normal model by embodying a distortion parameter. An application of the new test in interpreting the distribution of microarray gene expression data on blood stem cells is included for illustration purpose.

1. Introduction

Let $\phi(\cdot)$ and $\Phi(\cdot)$ be the pdf and cdf of the standard normal distribution. For any $\lambda \in R$, the skew normal model $SN(\lambda)$ refers to the random variable following the density function (Azzalini [2]):

$$f(x) = 2\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma}\right), \quad x \in R. \tag{1}$$

2000 Mathematics Subject Classification: Primary 62E15; Secondary 62P10.

Keywords and phrases: skew normal models, distorted normal distribution, moment generating function, likelihood ratio test, affymetrix data.

The research of Jie Chen is supported in part by the NSF grant DMS-0426148.

Received September 10, 2005

Notice that the normal model $X \sim N(\mu, \sigma)$ uses μ and σ to measure the mean and scale of the variability of the data, and $SN(\lambda)$ extends the $N(\mu, \sigma)$ model by introducing a new parameter λ. This λ elegantly allows the model to catch the asymmetric property of the data, and measures the skewness for the underlying distribution of the data. When $\lambda = 0$, (1) becomes the normal model.

For the extension of the skew normal model into multivariate scenario, Azzalini and Dalla Valle [4] discussed a multivariate version of the skew normal distribution, which was refined by another version of multivariate skew normal model that has coherency property (Gupta and Chen [11]). Branco and Dey [5] proposed a multivariate skew elliptical distribution. For a skew normal random sample, Chen et al. [7] derived the distribution of the sample mean.

The skew normal model has applications in many disciplines. For example, Azzalini and Capitanio [3] discussed the application of the multivariate skew normal model; Gupta and Chen [9] analyzed goodness-of-fit methods for fitting epidemiology data into the skew normal model; Chen et al. [8] applied the skew normal model to stock market data for the investigation of sell prices in a bull or (bear) market; and Kim and Mallick [12] discussed a method of Bayesian prediction using the skew normal model, among many others.

One of the remaining problems with the skew normal distribution in modeling is that the skew factor λ, in some cases may not be able to completely reflect the skewness conveyed by the variability of the data. Notice that in (1), when λ varies, $\Phi(\lambda x)$ is not sensitive to the change of x, especially when x is large enough. In other words, λ cannot reflect the shape of the tail for the distribution of the data. Under this scenario, it is natural to consider another parameter that can sensitively model the tail of the distribution of the data. This leads to the following model:

$$f(x) = c\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\lambda\frac{x-\mu}{\sigma} + b\right), \tag{2}$$

where the parameter b contained in model (2) serves as another dimension to model the data. In Section 2, we review the plausibility of

model (2). We then provide a probability interpretation or stochastic representation of the model in Section 3, which is then followed by Section 4 describing properties of the new model. Section 5 especially addresses a likelihood ratio test, which is applied to modeling microarray gene expression data on blood stem cell studies in Section 6.

2. Distorted Normal Model

In this section, we discuss a definition of the *distorted normal family*. In the sequel we denote $\phi(x, \sigma)$ and $\Phi(x, \sigma)$ the pdf and cdf of a normal random variable with mean 0 and variance σ^2, respectively.

Definition. If a continuous random variable X has the density

$$f(x) = c^* \phi(x, \sigma)\Phi(\lambda x + b, \sigma) \tag{3}$$

for any $\sigma > 0, b \in R, \lambda \in R$ and $c^* > 0$, then X is called a *distorted normal random variable* with skew factor λ and distortion factor b.

The following observation shows that the function $f(x)$ specified in the definition is eligible for being a density function when the constant, c^*, takes certain value.

Proposition 1. *Let* $f(x)$ *be the density of a distorted normal random variable defined in* (3), *when* $c^* = [\Phi(b, \sigma\sqrt{1+\lambda^2})]^{-1}$. *Then*

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{4}$$

The skew parameter λ and the distortion factor b in the model (3) measure the skewness of the shape of the density, and the quantity σ models the population variability.

Proposition 2. *When the distortion factor is set to zero* $(b = 0)$, *the distorted normal distribution becomes the skew normal distribution*:

$$f(x) = 2\frac{1}{\sigma}\phi(x/\sigma)\Phi(\lambda x/\sigma).$$

The proof of Proposition 2 is self-evident by the fact that $\Phi(0) = \frac{1}{2}$.

Specifically, when the parameter b is 0, for any $\sigma > 0$ the distorted normal distribution defined by (1) becomes the general skew normal distribution with pdf $f(x; \lambda, \sigma) = 2\phi(x, \sigma)\Phi(\lambda x, \sigma)$. When b is 0 and σ is 1, the distorted normal distribution becomes the skew normal distribution with pdf $f(x; \lambda) = 2\phi(x)\Phi(\lambda x)$, which has been extensively studied by Azzalini [2], Gupta and Chen [9, 10], among others. When both b and λ are set to 0, the distorted normal distribution defined in (1) becomes the normal distribution with mean 0 and variance σ^2. It is therefore evident that the distorted normal distribution family embraces the normal distribution family and the skew normal distribution family.

For convenience, we will denote the family of distorted normal distributions defined by (1) as $\mathcal{DSN}(b, \lambda, \sigma)$, and denote the skew normal distribution with pdf $f(x; \lambda) = 2\phi(x)\Phi(\lambda x)$ as $\mathcal{SN}(\lambda)$.

3. Probability Interpretation

In this section, we provide a probability interpretation for the model discussed in Section 2. Recall that the stochastic representation of the skew normal model is the following. If X and Y are two independent random variables following the standard normal distribution, then for any real number λ, the random variable $Z = \frac{\lambda}{\sqrt{1 + \lambda^2}} | X | + \frac{1}{\sqrt{1 + \lambda^2}} Y$ follows a skew normal distribution $SN(\lambda)$. In the following discussion, we start with two independent normal random variables to generate a distorted normal model.

Theorem 1. *Let U and V be two independent random variables, $U \sim N(0, \sigma_1^2)$ and $V \sim N(0, \sigma_2^2)$. Denote $Y = U - V$, then the density of Y when $U \geq c$ reads*

$$f(y) = k\phi(y; \sqrt{\sigma_1^2 + \sigma_2^2})\Phi\left(\frac{\sigma_1}{\sigma_2} y - \frac{c}{\sigma_1\sigma_2}(\sigma_1^2 + \sigma_2^2); \sqrt{\sigma_1^2 + \sigma_2^2}\right),$$

where c is the threshold of truncation for the random variable U, the coefficient $k = \left[\Phi\left(-\frac{c}{\sigma_1}, 1\right)\right]^{-1}$.

To prove Theorem 1, we need the following two results. First, notice that for constants $a > 0$, $b \in R$ and $\sigma > 0$,

$$\int_c^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left\{-\frac{1}{2\sigma^2}(ax - b)^2\right\} dx = \frac{1}{a} \Phi\left(b/\sigma - \frac{ac}{\sigma}\right). \tag{5}$$

Next, we shall prove a lemma that will play a key role in the proof of Theorem 1.

Lemma 1. *For positive constants* σ *and* σ_x,

$$\int_c^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-x)^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{x^2}{2\sigma_x^2}\right\} dx$$

$$= \phi(y; \sqrt{\sigma^2 + \sigma_x^2}) \Phi\left(\frac{\sigma_x}{\sigma} y - \frac{c}{\sigma\sigma_x}(\sigma^2 + \sigma_x^2); \sqrt{\sigma^2 + \sigma_x^2}\right), \tag{6}$$

where $\phi(y; \sqrt{\sigma^2 + \sigma_x^2})$ *and* $\Phi\left(\frac{\sigma_x}{\sigma} y - \frac{c}{\sigma\sigma_x}(\sigma^2 + \sigma_x^2); \sqrt{\sigma^2 + \sigma_x^2}\right)$ *are the pdf and cdf of the normal distribution with mean* 0 *and variance* $\sigma^2 + \sigma_x^2$ *at the points y and* $\frac{\sigma_x}{\sigma} y - \frac{c}{\sigma\sigma_x}(\sigma^2 + \sigma_x^2)$, *respectively.*

Proof. The left hand side of (6) reads

$$\Delta = \int_c^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-x)^2}{2\sigma^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{x^2}{2\sigma_x^2}\right\} dx$$

$$= \int_c^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right)^2 \frac{1}{\sigma\sigma_x} \exp\left\{-\frac{1}{2\sigma^2}(y^2 - 2xy + (\sigma_x^2 x^2 + \sigma^2 x^2)/\sigma_x^2)\right\} dx$$

$$= \int_c^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right)^2 \frac{1}{\sigma\sigma_x} \exp\left\{-\frac{1}{2\sigma^2}((\sigma^2 + \sigma_x^2)x^2/\sigma_x^2 - 2xy + \sigma_x^2 y^2/(\sigma^2 + \sigma_x^2))\right\}$$

$$\exp\left\{-\frac{1}{2\sigma^2} y^2\right\} \exp\left\{\frac{1}{2\sigma^2} \sigma_x^2 y^2/(\sigma^2 + \sigma_x^2)\right\} dx. \tag{7}$$

(7) can be written as

$$\Delta = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left\{-\frac{1}{2\sigma^2}(y^2 - \sigma_x^2 y^2/(\sigma^2+\sigma_x^2))\right\}$$

$$\int_c^\infty \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\sqrt{\sigma^2+\sigma_x^2}\,x/\sigma_x - \sigma_x y/\sqrt{\sigma^2+\sigma_x^2})^2\right\}dx. \tag{8}$$

By (5),

$$\int_c^\infty \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{1}{2\sigma^2}(\sqrt{\sigma^2+\sigma_x^2}\,x/\sigma_x - \sigma_x y/\sqrt{(\sigma^2+\sigma_x^2)})^2\right\}dx$$

$$= \frac{\sigma_x}{\sqrt{\sigma^2+\sigma_x^2}}\Phi\left(\frac{\sigma_x}{\sigma}y/\sqrt{\sigma^2+\sigma_x^2} - \frac{c}{\sigma\sigma_x}\sqrt{\sigma^2+\sigma_x^2}\right). \tag{9}$$

Thus putting (9) into (8) yields

$$\Delta = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma^2+\sigma_x^2}}\exp\left\{-\frac{1}{2}(y^2/(\sigma^2+\sigma_x^2))\right\}$$

$$\Phi\left(\frac{\sigma_x}{\sigma}y/\sqrt{\sigma^2+\sigma_x^2} - \frac{c}{\sigma\sigma_x}\sqrt{\sigma^2+\sigma_x^2}\right)$$

$$= \phi(y;\sqrt{\sigma^2+\sigma_x^2})\Phi\left(\frac{\sigma_x}{\sigma}y - \frac{c}{\sigma\sigma_x}(\sigma^2+\sigma_x^2);\sqrt{\sigma^2+\sigma_x^2}\right), \tag{10}$$

which is the right hand side of (6).

With the above lemma, we are able to derive Theorem 1 as follows.

Proof of Theorem 1. Notice that

$$P(Y \le y) = P(U - V \le y \,|\, U \ge c)$$

$$= k\int_c^\infty \int_{-\infty}^{y-x} \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_2}\exp\left\{-\frac{t^2}{2\sigma_2^2}\right\} f_X(x)\,dt\,dx$$

$$= k\int_c^\infty \int_{-\infty}^{y-x} \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_2}\exp\left\{-\frac{t^2}{2\sigma_2^2}\right\} \phi(x,\,\sigma_1)\,dt\,dx, \tag{11}$$

where $k = [P(U \ge c)]^{-1} = \left[\Phi\left(-\frac{c}{\sigma_1}\right)\right]^{-1} = \left[\Phi\left(-\frac{c}{\sigma_1}, 1\right)\right]^{-1}$.

From (11), the density of Y reads

$$f(y) = k\int_c^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_2} \exp\left\{-\frac{(y-x)^2}{2\sigma_2^2}\right\} \phi(x, \sigma_1)dx$$

$$= k\phi(y; \sqrt{\sigma_1^2 + \sigma_2^2})\Phi\left(\frac{\sigma_1}{\sigma_2} y - \frac{c}{\sigma_1\sigma_2}(\sigma_1^2 + \sigma_2^2); \sqrt{\sigma_1^2 + \sigma_2^2}\right),$$

by Lemma 1.

4. Properties of the Distorted Normal Distribution

We discuss various properties of $\mathcal{DSN}(b, \lambda, \sigma)$ in this section. To derive the moment generating function of $\mathcal{DSN}(b, \lambda, \sigma)$, we need the following result as in Zacks [15, pp. 53-54].

Lemma 2. *Let random variable* $Z \sim N(0,1)$. *Then*

$$E[\Phi(hZ + k)] = \Phi(k/\sqrt{1 + h^2}),$$

for any $h, k \in R$ *and* $\Phi(\cdot)$ *is the CDF of* $N(0, 1)$.

We can now state the second theorem.

Theorem 2. *Let random variable* $X \sim \mathcal{DSN}(b, \lambda, \sigma)$. *Then the moment generating function of X is given by*

$$M_X(t) = c^* e^{\sigma^2 t^2/2} \Phi\{(\lambda\sigma t + b/\sigma)/\sqrt{1 + \lambda^2}\} \tag{12}$$

for any $t \in R$ *with* c^* *as given in Proposition* 1.

Proof. With pdf given by (1), for any $t \in R$ the moment generating function of X is

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} c^* \phi(x, \sigma)\, \Phi(\lambda x + b, \sigma)dx$$

$$= c^* e^{\sigma^2 t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x - \sigma^2 t)^2}{2\sigma^2}\right\} \Phi(\lambda x + b, \sigma)dx$$

$$= c^* e^{\sigma^2 t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \Phi(\lambda y + \lambda\sigma t + b/\sigma)dy$$

$$= c^* e^{\sigma^2 t^2/2} \Phi\{(\lambda\sigma t + b/\sigma)/\sqrt{1 + \lambda^2}\},$$

in view of Lemma 2.

Using the moment generating function (12) of $\mathcal{DSN}(b, \lambda, \sigma)$, after some algebra, we obtain the first four non central moments of the distorted normal random variable X as follows:

$$E[X] = c^*\eta\delta\sigma,$$

$$E[X^2] = \sigma^2 - bc^*\eta\delta^2\sigma\xi,$$

$$E[X^3] = 3c^*\eta\delta\sigma^3 + b^2c^*\eta\delta^3\sigma - c^*\eta\delta^3\sigma^3,$$

$$E[X^4] = 3\sigma^4 - 3bc^*\eta\delta^2\sigma^3\xi - 3c^*\eta\delta\sigma^5 - b^3c^*\eta\delta^4\sigma^2b^3\xi + bc^*\eta\delta^4\sigma^2\xi + 2bc^*\eta\delta^4\sigma^3\xi, \tag{13}$$

where

$$\eta = \phi(b/[\sigma\sqrt{1+\lambda^2}]),$$

$$\xi = \frac{1}{\sqrt{1+\lambda^2}},$$

$$\delta = \lambda\xi. \tag{14}$$

Also, the variance of the distorted random variable X is obtained as:

$$Var[X] = \sigma^2(1 - c^{*2}\eta^2\delta^2) - c^*\eta\delta^2\xi.$$

Next we summarize some statistical properties of the distorted normal distribution $\mathcal{DSN}(b, \lambda, \sigma)$ in this section. As introduced in Section 2, the $\mathcal{DSN}(b, \lambda, \sigma)$ contains the skew normal distribution and normal distribution.

Theorem 3. *If* $X \sim \mathcal{DSN}(b, \lambda, \sigma)$, *then the following holds*:

(i) *When* $b = 0, \sigma = 1$, *then* $X \sim \mathcal{SN}(\lambda)$ *and* $X^2 \sim \chi_1^2$.

(ii) *When* $b = 0, \lambda = 0$, *then* $X \sim N(0, \sigma^2)$.

(iii) *As* $\lambda \to \infty$, $\mathcal{DSN}(b, \lambda, \sigma) \to 2\phi(x, \sigma)$, $x > 0$.

(iv) $X \sim \mathcal{DSN}(b, \lambda, \sigma)$, *then* $-X \sim \mathcal{DSN}(b, -\lambda, \sigma)$.

The proof of Theorem 3 is self-evident.

Theorem 4. *If* $X \sim \mathcal{DSN}(b, \lambda, \sigma)$, *and* $Z \sim N(0, \sigma^2)$, X *and* Z *are independent, then* $(X + Z)/\sqrt{2} \sim \mathcal{DSN}(b, \lambda^*, \sigma)$, *where* $\lambda^* = \delta/\sqrt{2}$.

Proof. Using the moment generating functions of $\mathcal{DSN}(b, \delta, \sigma)$ and $N(0, \sigma^2)$, the moment generating function of $(X + Z)/\sqrt{2}$ is

$$M_{(X+Z)/\sqrt{2}}(t) = M_X(t/\sqrt{2}) \cdot M_Z(t/\sqrt{2})$$

$$= c^* e^{\sigma^2 t^2/4} \Phi\{(\lambda\sigma t/\sqrt{2} + b/\sigma)/\sqrt{1 + \lambda^2}\} \cdot e^{\sigma^2 t^2/4}$$

$$= c^* e^{\sigma^2 t^2/2} \Phi\{(\lambda\sigma t + \sqrt{2}b/\sigma)/\sqrt{2(1 + \lambda^2)}\},$$

which is the MGF of $\mathcal{DSN}(b, \lambda^*, \sigma)$, with $\lambda^* = \delta/\sqrt{2}$.

5. Likelihood Ratio Test

Assume that $X_1, X_2, ..., X_n$ constitute a random sample. To test the hypothesis that the data is from a distorted normal with specific parameters b_1, λ_1, and σ_1, we have

$$H_0 : X_1, X_2, ..., X_n \sim iid\ N(0, \sigma^2),$$

versus the alternative hypothesis

$$H_1 : X_1, X_2, ..., X_n \sim iid\ \mathcal{DSN}(b_1, \lambda_1, \sigma).$$

Using (ii) of Theorem 3, testing the above null hypothesis H_0 versus the alternative hypothesis H_1 is equivalent to testing the null hypothesis

$$H_0 : \begin{pmatrix} b \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

versus the alternative hypothesis

$$H_1 : \begin{pmatrix} b \\ \lambda \end{pmatrix} = \begin{pmatrix} b_1 \\ \lambda_1 \end{pmatrix},$$

where b_1 and λ_1 are specified.

A test statistic for testing H_0 versus H_1 can be constructed using the likelihood ratio approach.

As the ratio of the likelihood functions under H_1 and H_0 is

$$\frac{f(X_1, X_2, ..., X_n; b_1, \lambda_1, \sigma)}{f(X_1, X_2, ..., X_n; b_0, \lambda_0, \sigma)} = \frac{\prod_{i=1}^{n} c^* \phi(X_i, \sigma)\Phi(\lambda_1 X_i + b_1, \sigma)}{\prod_{i=1}^{n} \phi(X_i, \sigma)}$$

$$\propto \prod_{i=1}^{n} \Phi(\lambda_1 X_i + b_1, \sigma).$$

The test statistic is based on $\prod_{i=1}^{n} \Phi(\lambda_1 X_i + b_1, \sigma)$ when σ is known. If σ is unknown, then we may use the sample standard deviation, which is a consistent estimator of σ in the test statistic. The rejection area is the set where $\prod_{i=1}^{n} \Phi(\lambda_1 X_i + b_1, \sigma)$ is greater than a constant, γ. Denote the order statistics of $X_1, X_2, ..., X_n$ as $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. For convenience, we consider the following two cases, separately.

Case (i). $\lambda_1 > 0$.

In this case, we have

$$\prod_{i=1}^{n} \Phi(\lambda_1 X_i + b_1, \sigma) \le \prod_{i=1}^{n} \Phi(\lambda_1 X_{(n)} + b_1, \sigma)$$

$$= \Phi^n(\lambda_1 X_{(n)} + b_1, \sigma).$$

We therefore construct an at most α-*level* test using $\Phi^n(\lambda_1 X_{(n)} + b_1, \sigma)$. When it is greater than some constant, then a decision of rejection of the null hypothesis H_0 can be made according a significance level α. Specifically, we reject H_0, if

$$\Phi(\lambda_1 X_{(n)} + b_1, \sigma) > \gamma,$$

or reject H_0, if

$$\Phi\left(\frac{\lambda_1 X_{(n)} + b_1}{\sigma}\right) > \gamma, \tag{15}$$

where γ is chosen such that

$$P\left[\Phi\left(\frac{\lambda_1 X_{(n)} + b_1}{\sigma}\right) > \gamma \,|\, H_0\right] = \alpha.$$

Observe that (15) is equivalent to

$$X_{(n)} > \frac{\sigma\Phi^{-1}(\gamma) - b_1}{\lambda_1},$$

where γ is such that

$$P\left[X_{(n)} > \frac{\sigma\Phi^{-1}(\gamma) - b_1}{\lambda_1} \,|\, H_0\right] = \alpha.$$

Using the CDF of the order statistics $X_{(n)}$ under H_0 and after some algebra $\Phi^{-1}(\gamma)$ is found to be

$$\Phi^{-1}(\gamma) = \frac{\lambda_1\sigma\Phi^{-1}[(1-\alpha)^{1/n}] + b_1}{\sigma}.$$

Therefore, we obtain the following decision rule:

$$\text{Reject } H_0, \text{ if } X_{(n)} > \sigma\Phi^{-1}[(1-\alpha)^{1/n}], \tag{16}$$

at the significance level up to α.

Case (ii). $\lambda_1 < 0$.

In this case, we first have

$$\prod_{i=1}^{n} \Phi(\lambda_1 X_i + b_1, \sigma) \le \prod_{i=1}^{n} \Phi(\lambda_1 X_{(1)} + b_1, \sigma)$$

$$= \Phi^n(\lambda_1 X_{(1)} + b_1, \sigma).$$

Similar to the development of case (i), we reject H_0, if

$$\Phi(\lambda_1 X_{(1)} + b_1, \sigma) > \tau,$$

where τ is such that

$$P\left[X_{(1)} < \frac{\sigma\Phi^{-1}(\tau) - b_1}{\lambda_1} \mid H_0\right] = \alpha.$$

Using the CDF of the order statistics $X_{(1)}$ under H_0 and after some algebra $\Phi^{-1}(\gamma)$ is found to be

$$\Phi^{-1}(\gamma) = \frac{\lambda_1\sigma\Phi^{-1}(\alpha^{1/n}) + b_1}{\sigma}.$$

Therefore, we obtain the following rule:

$$\text{Reject } H_0, \text{ if } X_{(1)} < \sigma\Phi^{-1}(\alpha^{1/n}), \tag{17}$$

at the significance level up to α.

In summary, (16) specifies the rejection rule for the case of the given $\lambda_1 > 0$ at the significance level up to α; and (17) gives the rejection rule for the case of $\lambda_1 < 0$ at the significance level up to α. In all cases, if σ is unknown, it is replaced by the sample standard deviation S for large sample size. In what follows, we will provide an example in accessing the distorted normal feature of a gene expression data set.

6. Application: Modeling an Affymetrix Data Set

Recent advances in biomedical technology result in one of the predominant devices – the microarray gene chip. Using a microarray chip, biologists can simultaneously obtain thousands or tens of thousands of numerical readings (gene expressions) for genes under investigation. These gene expressions are obtained by hybridizing target experimental units (containing abundant mRNA's) with the probe sets pre-specified on the microarray chip. The hybridization intensity at each gene location is then read through a special software after normalization, giving the

numerical expression for each gene. Using the expressions of the thousands of genes in a living cell, biologist can grasp the numerical aspects of the genes through gene profiling. One of the commercially available arrays is the affymetrix microarray gene chip. The affymetrix gene chip probe array provides an average difference (an expression index) for each gene. That is, the average difference serves as an indicator for the level of gene expression. It is then applied to determine the change in the hybridization intensity of a given probe set. For affymetrix gene chips, the reading (average difference in gene expression) is calculated using the sum of the "perfect match-mismatch" for each probe pair in a probe set divided by the number of probe pairs used in the probe set. That is

$$AvDiff = \frac{1}{|A|}\sum_{j\in A}(PM_j - MM_j),$$

where A is the subset of probes for which the ranges of $PM_j - MM_j$ are within 3 SDs away from the average of $PM_{(2)} - MM_{(2)}$, ..., $PM_{(J-1)} - MM_{(J-1)}$ with J being the number of probe sets used in the array, $|A|$ is the cardinality of set A, and $PM_{(j)} - MM_{(j)}$ is the j-th smallest difference. The average difference is given when processing the hybridized experimental sample on the microarray chip via the software of affymetrix gene chip data mode. It is observed that the affymetrix gene expression (average difference) data for an experiment is usually not symmetric. Could it be distorted normal?

To seek a legitimate assumption for the truncated model, we start with the general assumption in Li and Wang [13] that the background adjusted and normalized gene expression data follow a normal model. Under the assumption that observations of a perfect match reading U and a mismatch reading V follow two independent normal distributions, we can apply the probability model discussed in Section 3 into this scenario. Since the reading of PM is discarded or replaced by certain threshold value when the reading is too low (for example, Akashi et al. [1]), we consider the distribution of $Y = U - V \,|\, U \geq c$ for a pre-specified constant c.

Now, according to Theorem 1 (with U and V being the perfect-match and mismatch readings, respectively), the density of average differences in the affymetrix gene expression data $Y = U - V \,|\, U \geq c$ is a distorted normal distribution defined in (1) with mean 0, variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$, skew factor $\lambda = \frac{\sigma_1}{\sigma_2}$, and distortion parameter $b = -\frac{c}{\sigma_1\sigma_2}(\sigma_1^2 + \sigma_2^2)$, where

$$\Phi(b, \sigma\sqrt{1+\lambda^2}) = \Phi\left(-\frac{c}{\sigma_1}\right).$$

Therefore, when the data of perfect match probe pairs are truncated according to certain value, the underlying distribution of the truncated data turns out to be a distorted normal distribution rather than being a normal or log normal distribution. This conclusion can be further verified by using the inference procedure proposed in Section 5 when a microarray data set is available.

A microarray gene expression data on hematopoietic stem cell (HSC) was obtained in Akashi et al. [1] for studying the development of blood stem cells. There are 5253 gene expressions passed through initial screening for further biological study. A histogram of the 5253 gene expression data in HSC shows a clear skew trend. To investigate whether the distribution of the gene expressions in HSC is normal or distorted normal, as an example, we test the null hypothesis

$$H_0 : \begin{pmatrix} b \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

versus the alternative hypothesis

$$H_1 : \begin{pmatrix} b \\ \lambda \end{pmatrix} = \begin{pmatrix} 208 \\ 5 \end{pmatrix},$$

where the value of $b_1 = 208$ and $\lambda = 5$ are pre-specified (see [1]). Calculation shows that for the sample size $n = 5253$ and significance level $\alpha = 5\%$, $\Phi^{-1}[(1-\alpha)^{1/n}] = 4.2702$. Also the sample standard deviation

is obtained as $\hat{\sigma} = 1158.9$. Then the right hand side of (16) is

$$\sigma\Phi^{-1}[(1-\alpha)^{1/n}] = S\Phi^{-1}[(1-\alpha)^{1/n}]$$

$$= 1158.9(4.2702)$$

$$= 4948.7347.$$

Since the test statistic value of $X_{(n)}$ is 19434.9, that is, $X_{(n)} > \sigma\Phi^{-1}[(1-\alpha)^{1/n}]$, we then conclude that H_0 should be rejected according to (16) at significance level $\alpha = 5\%$.

References

[1] K. Akashi, X. He, Jie Chen, H. Iwasaki, C. Niu, B. Steenhard, J. Zhang, R. Perera, J. Haug and L. Li, Transcriptional accessibility for multi-tissue and multi-hematopoietic lineage genes is hierarchically controlled during early hematopoiesis, Blood 101(2) (2003), 383-390.

[2] A. Azzalini, A class of distribution which includes the normal ones, Scand. J. Statist. 12 (1985), 171-178.

[3] A. Azzalini and A. Capitanio, Statistical applications of the multivariate skew normal distribution, J. R. Stat. Soc. Ser. B Stat. Methodol. 61(3) (1999), 579-602.

[4] A. Azzalini and A. Dalla Valle, The multivariate skew-normal distribution, Biometrika 83 (1996), 715-726.

[5] M. C. Branco and D. K. Dey, A general class of multivariate skew-elliptical distributions, J. Multivariate Anal. 79(1) (2001), 99-113.

[6] J. T. Chen, Jie Chen and A. Gupta, A statistical model for micro-array gene expression data, Department of Mathematics and Statistics, BGSU. No. 03-03, 2003.

[7] J. T. Chen, A. K. Gupta and T. T. Nguyen, The density of the skew normal sample mean and its applications, J. Stat. Comput. Simul. 74(7) (2004), 487-494.

[8] J. T. Chen, A. K. Gupta and C. Troskie, Distribution of stock returns when the market is up (down), Comm. Statist. Theory Methods 32 (2003), 1541-1558.

[9] A. Gupta and T. Chen, Goodness-of-fit tests for the skew-normal distribution, Comm. Statist. Simnulation Comput. 30(4) (2001), 907-930.

[10] A. K. Gupta and T. Chen, On the sample characterization criterion for normal distributions, J. Stat. Comput. Simul. 73 (2003), 155-163.

[11] A. K. Gupta and J. T. Chen, A class of multivariate skew normal models, Ann. Inst. Statist. Math. 56(2) (2004), 305-315.

[12] H. M. Kim and B. K. Mallick, A Bayesian prediction using the skew Gaussian distribution, J. Statist. Plann. Inference 120 (2004), 85-101.

[13] C. Li and W. Wang, Model based analysis of nucleotide arrays; expression index computation and outlier detection, Proc. Natl. Acad. Sci. USA 98 (2001), 31-36.

[14] P. Tamayo, A. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander and T. R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. 96 (1999), 2907-2912.

[15] S. Zacks, Parametric Statistical Inference, Pergamon Press, Oxford, 1981.

■