Journal of Advanced Statistics, Vol. 1, No. 4, December 2016
https://dx.doi.org/10.22606/jas.2016.14003

A Generalized Mixture Model for Detecting Differentially Expressed Genes in Microarray Experiments

Mehdi Razzaghi* and Dong Zhang

Mathematical and Digital Sciences, Bloomsburg University in Pennsylvania, 400 East 2nd Street, Bloomsburg, PA, United States
Email: : mrazzagh@bloomu.edu

Abstract To determine the genes that are differentially expressed between samples in microarray experiments, traditionally the expression levels were assessed by taking the intensity levels at a spot on the array and flagging the gene if the magnitude of the fold change exceeded a threshold. Recently, however, there has been much effort to improve the methodology by incorporating the variability of the intensity ratios. While the Student's t-test and several of its variants have been proposed by several authors, a methodology that has found widespread popularity is the application of the mixture model in a hierarchical approach whereby the mean of the distribution of the normalized log ratios is assumed to be a random variable having a mixture of two components. One component is a point mass distribution concentrated at zero to represent the non-differentially expressed genes and another component is a suitable distribution with zero mean to represent the differentially expressed genes. The normal and the Laplace distributions have been previously suggested for the differentially expressed genes component of the mixture. But, once again, the symmetry assumption can make these distributions unsuitable. Here, we take a more general approach and apply the beta-normal model to describe the distribution of the mean of the differentially expressed genes. The advantage of this approach is that we no longer assume symmetry for the distribution and let the data determine its shape. We show that our approach includes the earlier results based on normality assumption as a special case. Simulation results demonstrate that there are advantages in using the more general beta-normal distribution. An example with a microarray experimental data is utilized to provide further illustration.

Keywords: Microarray experiments, differential expression, beta-normal, posterior odds

1 Introduction

In recent years, microarray experiments have received an increasing amount of attention in scientific research. This new technology allows the scientist to simultaneously examine the expression levels of thousands of genes. Changes in the expression levels across multiple samples provide useful information that can be helpful in generating hypotheses relating gene expression levels to other characteristics. Consequently, microarray experiments have become an important tool for investigation of the molecular basis of many diseases. In most cases, the experiment is repeated under two different conditions such as treatment versus control with the goal of identifying the genes that are differentially expressed under the two conditions. A challenging statistical problem that has evolved in this area has been the development of a methodology to detect a statistically significant difference in the intensity levels of genes. The oldest and the simplest method for identification of differential expression in genes is the 'Fold Change' method ([1]), where two genes are declared as being differentially expressed if the log intensity levels differ by more than a fixed, but arbitrary value. Thus if for example the cut-off value is selected to be a two-fold difference, then genes under the two experimental conditions are differentially expressed if the expression level of one is at least twice as the other. The fold change method has received much criticism from the statistical community since it does not account for variability among genes and thus does not have any statistical properties. Moreover, if the data are not appropriately normalized, the fold change method is subject to severe bias. The t-test, on the other hand, has been considered as a simple alternative to the fold change method since it accounts for variability and has well known statistical properties ([2]). In replicated experiments, one can estimate the variance for each gene and use it in the calculation of the t-score. A major problem, however, is that with only a small number of replicates, the variance can be

grossly underestimated, leading to erroneous results and large type I error. An improved estimate of the variance may be obtained if the variability across all genes is used. But, that would entail the assumption of equal variance across genes, and by ignoring variability across genes, the t-test will effectively be equivalent to the fold change method ([3]).

A modification of the t-test called the S-test was introduced by [4], where the authors suggested adding a small positive constant penalty factor to the denominator of the gene-specific t-test. This penalty factor was to be selected as the 90th percentile of the gene specific standard deviations. In what became known as the SAM (Significance Analysis of Microarray) test, [5] proposed choosing the penalty factor so that the coefficient of variation of the absolute values of the t scores is minimized. [6] introduced the use of distribution mixtures for analyzing the microarray gene expression data and since then, a large number of papers have considered different mixture models for various aspects of statistical analysis of microarray data. One approach is to use a statistical test of hypothesis such as the t-test or the SAM test for each gene and model the p-values as a mixture of a uniform and Beta distributions. This approach was adopted by [7]. Another approach is to assume a hierarchical structure in which the log intensity ratio is modeled as a Normal distribution for which the mean is assumed to follow a mixture model. The B test, introduced by [8] proposed the use of a Normal mixture for detecting differentially expressed genes. The test uses the log posterior odds to rank the genes. The method combines information across many genes and is proved to be more stable than the t-test. [9] noted that the assumption of normality for the components of the mixture is quite strong and in many cases not realistic. They suggested the use of a Laplace distribution as a long-tailed alternative to the Normal and modeled the mean relative expression levels as a Laplace mixture. [10] considered the numerical comparison of the efficiency of several gene selection criteria. More recently, [11] also note that the assumption of normality for the distribution of the gene expression levels may not be true and moreover, all the genes may not have a common distribution. They suggest the use of a generalized logistic distribution of type II whereby the genes are partitioned into several groups based on their expression levels. A test is then derived for detecting genes with differential expression levels. Here, we explore the use of the Beta-Normal distribution to model the gene expression data. Following [8] and [9], we also take a distribution mixture approach. But, where these authors assume normality or the Laplace distribution to express the distribution of the mean of the differentially expressed genes, we use a Beta-Normal. We believe that due to the generality and flexibility of this model, a more realistic interpretation of the data may be obtained. The advantage of our approach is that we let the data determine the shape and the level of skewness that best describes the expression levels of the differentially expressed genes. The properties of the Beta-Normal distribution will be briefly described in the next section. In Section 3, we describe our modeling procedure and discuss the parameter estimation in Section 4. Section 5 is devoted to a simulation study to demonstrate our methodology and compare it to similar methods. In Section 6, an example using a data set will be presented for illustration.

2 Beta-Normal Distribution

For a cumulative distribution function $G(.)$, a new general class of distributions can be defined as

$$F(x) = \frac{1}{B(\alpha,\beta)} \int_0^{G(x)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta, \alpha > 0, \beta > 0, \tag{1}$$

where $B(\alpha,\beta)$ is the Beta function given by

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

and

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt.$$

[12] used a normal distribution with mean μ and standard deviation σ for $G(.)$ and called the resulting distribution the Beta-Normal distribution. Thus, from (1), the density of the Beta-Normal distribution is given by

$$f(x;\mu,\sigma) = \frac{1}{\sigma B(\alpha,\beta)} \left\{\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{\alpha-1} \left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{\beta-1} \phi\left(\frac{x-\mu}{\sigma}\right) \tag{2}$$

https://dx.doi.org/10.22606/jas.2016.14003

where $\Phi(.)$ and $\phi(.)$ are respectively the CDF and the PDF of a standard Normal distribution. Note that in the special case $\alpha = \beta = 1$, the normal distribution is resulted. An equivalent representation of the Beta-Normal distribution is as follows. Define the random variable X as

$$X = \Phi^{-1}(u)$$

where u is a number between 0 and 1. It is well-known that if u is the realized value of a uniform $U(0,1)$ distribution, then X is a standard Normal variable . Now, more generally, if u is the realization of a Beta random variable with parameters α and β, then the resulting random variable X has a Beta-Normal distribution, where, as before, in the special case $\alpha = \beta = 1$, normality is derived. This latter representation is particularly useful for simulation of Beta-Normal variables. [12] discuss several interesting properties of the Beta-Normal distribution some of which are worth mentioning here. First, the distribution is symmetric about μ when α=β, becoming positively skewed when $\alpha > \beta$ and negatively skewed when $\alpha < \beta$. This property is particularly useful for modeling gene expression levels since in practice, estimates of α and β can determine the type and degree of skewness in the fitted distribution. The mean of the Beta-Normal distribution is an increasing function of α and a decreasing function of β while the standard deviation of the distribution is a decreasing function of both α and β. The kurtosis is an increasing function of α for a fixed $\beta < \alpha$ and a decreasing function of β for a fixed $\alpha < \beta$. [12] also derive a closed form expression for the first moment of the Beta-Normal distribution and evaluate it for some specific integer values of α and β. [13] also consider the moments of Beta-Normal distribution and derive a general expression for moments of the distribution when α and β are integers. The bimodality property of the Beta-Normal distribution is discussed in [14]. They show that the distribution becomes bimodal for certain values of parameters α and β. While analytical solutions for values of α and β for which bimodality occurs cannot be obtained, the authors show numerically that generally, bimodality occurs for values of α and β below 0.214. Interestingly, however, the modality of the distribution is independent of the parameters μ and σ. Note also that when $\alpha = 2$ and $\beta = 1$, the skew-normal distribution (Azzalini, 1985) with shape parameter of 1 is resulted. Beta-Normal distribution has also been successfully used in some applications. For example, [15] utilized this distribution to model dose-response data and estimate the risk in toxicological experiments. A density function plot using different choices of α and β with same $\mu = 2$ and $\sigma = 1$ can be seen in Figure 1.

3 Modeling Mean Expression Levels

Let $X_{ij}, i = 1, \ldots, N, j = 1, \ldots, n_i$ be the normalized log ratio of the expression levels in the two conditions for gene i with n_i replications. Assume that X_{ij} is a random variable from a Normal distribution with mean μ_i and variance σ_i^2 for $i = 1, \ldots, N$. Following [8], for the gene i we define an indicator random variable I_i which is equal to 1 if the gene is differentially expressed under the two conditions and takes the value 0 if the gene is not differentially expressed. If a gene is differentially expressed, then μ_i is assumed to have a distribution with mean 0. Otherwise, if the gene is not differentially expressed, its mean expression level is zero. Hence, if p is the proportion of differentially expressed genes, then μ_i can be regarded as a random variable having a mixture of two distributions $pf(\mu_i; 0, \sqrt{v}\sigma_i) + (1-p)\delta(0)$ where $\delta(0)$ denotes the point-mass density with mass at 0 and $f(\mu_i; 0, \sqrt{v}\sigma_i)$ is a density function with mean 0 and standard deviation $\sqrt{v}\sigma_i$, where v is a hyper-parameter expressing the dependence between the priors for μ_i and σ_i. Now, [8] propose the use of Normal density for $f(\mu_i; 0, \sqrt{v}\sigma_i)$. Noting that the Gaussian variation is an exception rather than the rule in practice, [9] suggest the use of Laplace distribution as a long tailed alternative to normality. Here, we use the Beta-Normal family introduced in the previous section to describe the distribution of μ_i. Thus if we let the prior distribution of $\tau_i = \sigma_i^{-2}$ be gamma,

$$g(\tau_i) = \frac{1}{\Gamma(b)} a^b \tau_i^{b-1} \exp\{-a\tau_i\}. \tag{3}$$

Then, we have

$$f_{I_i=0}(\mu_i) = \delta(0),$$

and

$$f_{I_i=1}(\mu_i|\tau_i) = \frac{\sqrt{\tau_i}}{\sqrt{v}B(\alpha,\beta)} \left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1} \left\{1 - \Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1} \phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right),$$

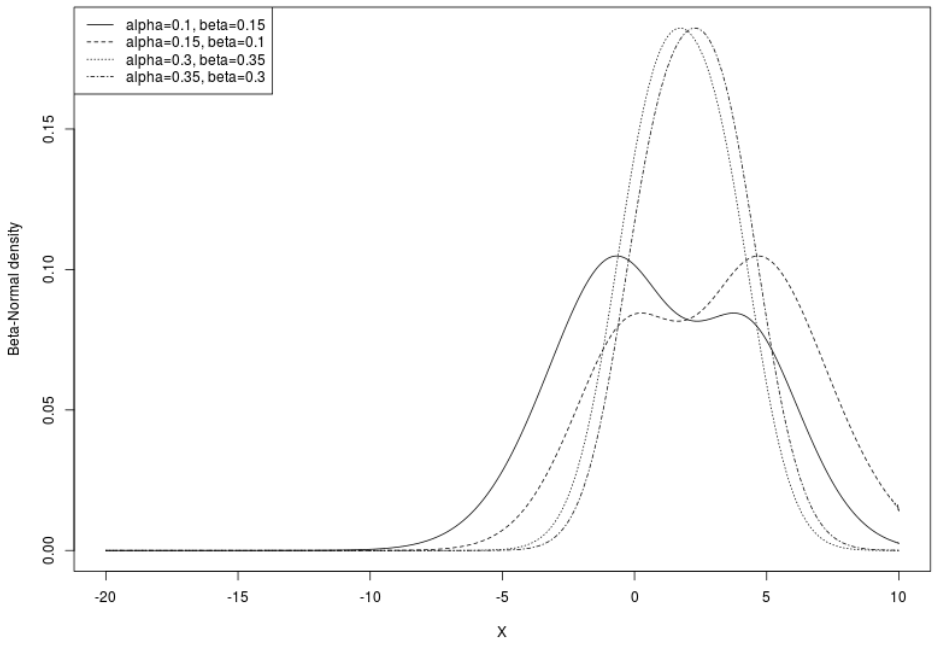


Figure 1. A density plot of Beta-Normal distribution using different values of α and β with $\mu = 2$ and $\sigma = 1$

where $\delta(.)$ is the Dirac's delta function, for $i = 1, 2, \ldots, N$. Applying the empirical Bayes method approach to calculate the log posterior odds, we have

$$BN_i = \log \hat{a}A\c{a} \left\{ \frac{P(I_i = 1 \big| X_i)}{P(I_i = 0 \big| X_i)} \right\} = \log \hat{a}A\c{a} \left\{ \frac{p}{1-p} \frac{f_{I_i=1}(X_i)}{f_{I_i=0}(X_i)} \right\} \tag{4}$$

as the test statistic, where $X_i = (X_{i1}, \ldots, X_{in_i})$. We show in Appendix 1 that the statistic BN_i can be written as

$$BN_i = \log \hat{a}A\c{a} \left(\frac{p}{1-p} \frac{\left[a + \frac{n_i}{2}(S_i^2 + \bar{X}_i^2)\right]^{b+\frac{n_i}{2}}}{\sqrt{2\pi v} B(\alpha, \beta) \Gamma(b + \frac{n_i}{2})} \times \right.$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}} \exp\left\{ -\tau_i \left[a + \frac{n_i}{2} S_i^2 + \frac{n_i \bar{X}_i^2}{2(1+vn_i)} \right] \right\} \int_{-\infty}^\infty \left\{ \Phi\left(\frac{\mu_i \sqrt{\tau_i}}{\sqrt{v}} \right) \right\}^{\alpha-1} \left\{ 1 - \Phi\left(\frac{\mu_i \sqrt{\tau_i}}{\sqrt{v}} \right) \right\}^{\beta-1} \times$$
$$\left. \exp\left\{ -\frac{(1+vn_i)\tau_i}{2v} \left[\mu_i - \frac{vn_i \bar{X}_i^2}{1+vn_i} \right]^2 \right\} d\mu_i d\tau_i \right) \tag{5}$$

for $i = 1, \ldots, N$, where $S_i^2 = 1/n_i \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ and $\bar{X}_i = 1/n_i \sum_{j=1}^{n_i} X_{ij}$. Note that (5) is a direct generalization of the statistic B derived by [8]. Indeed, in Appendix 2 we show in the special case when α and β are both equal to 1, BN, given by (5), collapses to the B statistic. Thus the extra parameters make BN much more flexible and we believe more suitable since we are getting away from the normality assumption of the data.

4 Parameter Estimation

Suppose that for each gene i, n_i, $i = 1, \ldots, N$ is the number of replications of the experiment. Then, if $X_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, N$ denotes the normalized log ratio of the expression levels for the j-th

https://dx.doi.org/10.22606/jas.2016.14003

replication of the i-th gene, we have

$$f(X_{i1},\ldots,X_{in_i}|p,v,\alpha,\beta,a,b)=\int_{\infty}^{\infty}\int_{0}^{\infty} f(X_{i1},\ldots,X_{in_i}|\mu_i,\tau_i)f(\mu_i|\tau_i)g(\tau_i)d\tau_i d\mu_i, \tag{6}$$

where in (6),

$$f(X_{i1},\ldots,X_{in_i}|\mu_i,\tau_i)=\left(\frac{\tau_i}{2\pi}\right)^{-\frac{n_i}{2}}\exp\left\{-\sum_{j=1}^{n_i}\frac{\tau_i(x_{ij}-\mu_i)^2}{2}\right\},$$

and

$$f(\mu_i|\tau_i)=\frac{p\sqrt{\tau_i}}{\sqrt{v}B(\alpha,\beta)}\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}\phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)+(1-p)\delta(0).$$

The log-likelihood can be expressed as

$$L(p,v,\alpha,\beta,a,b)=\sum_{i=1}^{N}\log f(X_{i1},\ldots,X_{in_i}|p,v,\alpha,\beta,a,b), \tag{7}$$

and the maximum likelihood estimates (MLEs) of the parameters can be obtained by maximizing (7). Clearly, no closed form solution for the likelihood equations exists and in order to derive the MLEs, we use a global optimization function based on the quasi-Newton method, mle() in the package stats4 (https://stat.ethz.ch/R-manual/R-devel/library/stats4/html/00Index.html) in R. More details about the quasi-Newton method can be found in [16]. The advantage of using the global optimization subroutine instead of solving the likelihood equation system is that the latter involves solving more integrals numerically which in turn leads to lower efficiency. In addition, we tried to use Markov Chain Monte Carlo (MCMC) approach for the numerical integration instead of the optimization subroutine in R. The idea improved the calculation efficiency and gave fast response, yet the accuracy was reduced.

5 Simulation

In this section, we use simulation to compare our proposed method with the Laplace model proposed by [9]. We mimic their simulation study by generating 100 data sets each with $N = 3000$ genes with two ($n = 2$) replicates from the Beta-Normal mixture distribution with the mixing proportion of $p = 0.1$ and $v = 1.2, 2$ and 3. To generate the variances σ_i^2 the reciprocal of gamma random variable τ_i generated from a gamma distribution (3) with parameter values $a = 0.04$ and $b = 2.8$ are utilized. A representative simulated dataset can be viewed in Figure 2. The summary statistics for the model parameter estimates are given in Table 1.

Table 1. The root mean squared error × 100 (bias × 100) of parameters

	p	v	α	β	a	b
v=1.2	1.312(-0.093)	1.860(1.194)	4.320(-0.010)	6.780(0.301)	8.897(0.629)	5.957(-1.772)
v=2	2.602(-0.077)	10.950(1.734)	3.993(0.042)	6.516(0.278)	9.319(1.994)	7.872(-1.709)
v=3	2.602(-0.082)	7.293(2.629)	3.291(0.105)	6.373(0.263)	3.367(2.486)	10.277(-1.661)

Next, we generated 100 data sets each with $N = 3000$ genes with $n = 2, 4, 6$ replicates from the Student t distribution as the general case of a symmetric distribution assumption. We employed degrees of freedom of $2/(1 - \tau_i)$, where τ_i is generated from the gamma distribution (3) with parameter values $a = 0.04$ and $b = 2.8$. The results of the comparison between Beta-Normal model and Laplace model are summarized in Table 2. We can conclude from Table 2 that both Beta-Normal mixture model and

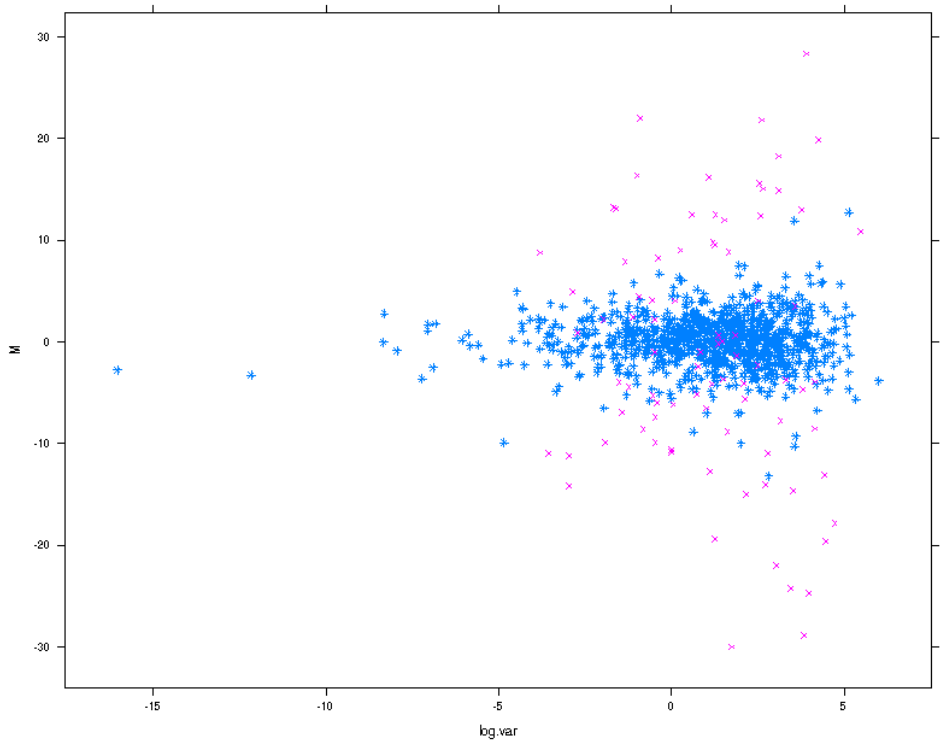


Figure 2. $\log_2$ fold change versus log variance plot of a simulated data set

Laplace mixture model yield large biases, yet the rMSEs of Beta-Normal model are uniformly smaller than Laplace model on estimating the mixture proportion. In the estimate of distribution parameters, a and b, of the inverse gamma distribution, the biases are almost the same for Beta-Normal model and Laplace model. Yet the Beta-Normal model still yields relatively smaller rMSEs.

Table 2. The root mean squared error ×100 (bias ×100) of parameters using difference replications

Replication n=2	p	a	b
Beta-Normal	0.450 (-0.970)	125.371 (1.523)	2.883 (-0.933)
Laplace	18.112 (-0.948)	127.990 (1.527)	2.891 (-0.933)
Replication n=4			
Beta-Normal	0.322 (-0.970)	56.858 (1.626)	1.390 (-0.937)
Laplace	15.749 (-0.842)	57.663 (1.631)	1.400 (-0.937)
Replication n=6			
Beta-Normal	0.313 (-0.915)	48.555 (1.597)	0.944 (-0.880)
Laplace	15.782 (-0.854)	47.316 (1.816)	0.938 (-0.990)

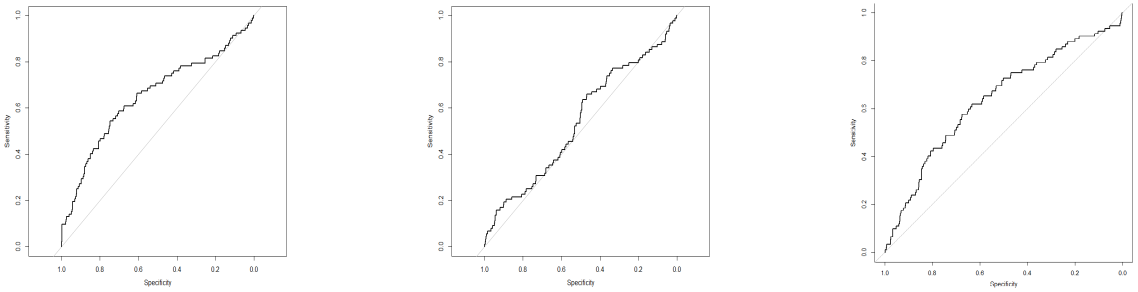
As pointed out earlier, our BN statistic is a generalization of the B statistic derived by [8] and in the special case that both α and β are equal to 1, B statistic is resulted. We therefore can assess the accuracy of our proposed BN statistic based on the Beta-Normal mixture model with both the B statistic of [8] using a normal mixture and the ρ statistic of [9]. The receiver operating characteristic (ROC) curve was employed for the comparison and the summary measurement of accuracy. The area under the ROC curve (AUC), was computed in the following three scenarios:

https://dx.doi.org/10.22606/jas.2016.14003

a. Assume $\mu_i|\tau_i$ is normally distributed. Figure 3 shows ROC curves of Laplace mixture, Beta-Normal mixture model and Normal mixture respectively.

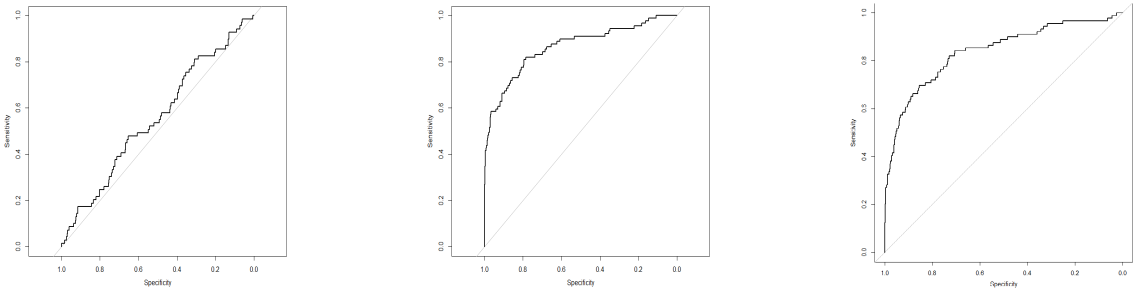
b. Assume $\mu_i|\tau_i$ is a Beta-Normal random variable. ROC curves can be found in Figure 4 for Laplace mixture, Beta-Normal mixture and Normal mixture models respectively.

c. Assume $\mu_i|\tau_i$ is a shifted gamma random variable, with shift parameter of 10. Figure 5 demonstrates the ROC curves for the three models.



(a) Laplace mixture model (b) Beta-Normal mixture model (c) Normal mixture model

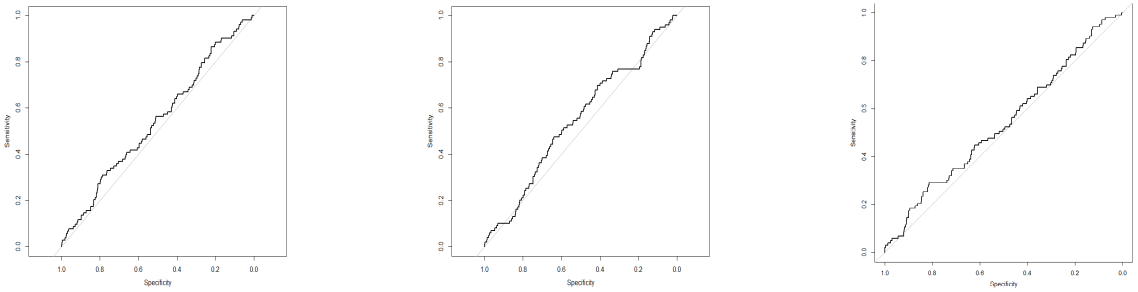
Figure 3. Comparison under Normal assumption.



(a) Laplace mixture model (b) Beta-Normal mixture model (c) Normal mixture model

Figure 4. Comparison under Beta-Normal assumption.

The AUC values are summarized in the Table 3. We found that in Scenario a, the accuracy of Beta-Normal fit was about 10% off the Laplace model and Normal model. Thus when the normality is attained, Normal mixture and Laplace mixture models may be preferred. Under Scenario b, the accuracy of Beta-Normal model was about 30% higher than Laplace model and about 3% higher than Normal model. Therefore in cases where normality assumption ceases to be true, which according to [11] is more realistic, the beta-normal mixture model performed the best. Using Scenario c, we noticed that the accuracy of Beta-Normal model is about 5.18% higher than Laplace model and 5.77% more than Normal model. Therefore, Beta-Normal mixture model shows better accuracy in fitting the model when normality is violated.



(a) Laplace mixture model (b) Beta-Normal mixture model (c) Normal mixture model

Figure 5. Comparison under gamma assumption.

Table 3. AUC values of three scenarios

	Laplace model	Beta-Normal model	Normal model
Normal assumption	0.6490	0.5365	0.6374
Beta-Normal assumption	0.5544	0.8631	0.8382
gamma assumption	0.5314	0.5832	0.5255

6 Example

We will demonstrate the proposed method using data publicly available from the Stanford Microarray Database (SMD, http://genome-www5.stanford.edu/). The experiments were carried out to compare the general effect on disease resistance RNA transcript levels of Arabidopsis thaliana infected by rhizobacterium Pseudomonas thivervalensis (strain MLG45) to axenic control plants ([17]). Here we consider the experiment on plant leaves, which has slides containing 16416 spots for 4 biological replicates hybridized to a common reference (SMD Experiment ID numbers 27084, 27000, 26995, and 26718). Further details are available in [17]. From these normalized log ratios, we estimated the parameters for the Beta-Normal model in Table 4. In [9], the estimate of the proportion of differentially expressed gene p is 0.075 and the parameters of the inverse gamma distribution are 5.89 and 3.39; meanwhile we found the corresponding values to be 0.06 for p and 2.17 and 3.61 for the inverse gamma parameters. We may conclude the estimates are close, however, our model estimates a lower proportion of differentially expressed genes.

Table 4. The estimates of Beta-Normal parameters in the Example.

p	v	α	β	a	b
0.06	0.01	10.0	8.74	2.17	3.61

7 Discussion

Traditionally, a common practice in the analysis of microarray data has been to assume that the distribution of the underlying ratio of the log intensity levels is normal. Several authors have shown that this assumption does not necessarily hold true and that even after some appropriate transformation and preprocessing, the distribution of the expression levels may still be non-normal, see for example [18], [19] and [20]. For this reason, the use of two-component distribution mixtures whereby one component represents the differentially expressed genes and the other represents the non-differentially expressed

gene has found popularity. By the same token, however, we believe that assuming a symmetric model to express the distribution of the differentially expressed genes may not be appropriate. The use of the family of Beta-Normal models to represent the distribution of the mean of the differentially expressed genes appears to have many advantages. This two parameter family encompasses a wide range of shapes including non-symmetric and bimodal distributions. Therefore the generality and versatility of the model makes it an attractive candidate for the proposed application. We believe that the fact that our BN statistic reduces to the normal model of [8] when both parameters α and β are equal to 1, is a significant improvement and makes the Beta-Normal model very appealing. It is surely more natural to let the data determine the extent of skewness and shape of the distribution. The two parameters α and β of the Beta-Normal distribution, estimated from the data provide this information about the underlying distribution. In fact from 1, the results of simulation show that in practice it is possible that the values of α and β be quite different from unity. The parameter estimates derived for the data in the illustrative example also indicate that the values of these parameters may be different from one, further proving that the distribution of the differentially expressed genes may be quite non-normal.

References

1. M. Schena, D. Shalon, R. Davis, and P. Brown, "Quantitative monitoring of gene expression patterns in complementary dna microarray," *Science*, vol. 270, pp. 467–470, 1995.
2. M. Callow, S. Dudoit, E. Gong, T. Speed, and E. Rubin, "Microarray expression profiling identifies genes with altered expressionin hdl-deficient mice," *Genome Research*, vol. 10, pp. 2022–2029, 2000.
3. X. Cui and G. Churchill, "Statistical tests for differential expression in edna microarray experiments," *Genome Biology*, vol. 4, p. 210, 2003.
4. B. Efron, R. Tibshrani, V. Goss, and G. Chu, *Microarrays and their use in a comparative experiment, Technical Report*. Department of Health Research and Policy, Stanford University, Stanford, CA, 2000.
5. V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 1, pp. 5116–5121, 2001.
6. B. Allison, G. Gadbury, M. Heo, J. Fernandez, C. Lee, Y. Prolla, and R. Weindruch, "A mixture model approach for the analysis of microarray gene expression data," *Computational Statistics and Data Analysis*, vol. 39, pp. 1–20, 2002.
7. R. Delongchamp, T. Lee, and C. Velasco, "A method for computing the overall statistical significance of a treatment effect among a group of genes," *BMC Bioinformatics*, vol. 11, 2006.
8. I. Lonnstedt and T. Speed, "Replicated microarray data." *Statistica Sinica*, vol. 12, pp. 31–46, 2002.
9. D. Bhowmick, A. Davison, D. Goldstein, and Y. Ruffieux, "A laplace mixture model for identification of differential expression in microarray experiments," *Biostatistics*, vol. 7, pp. 630–641, 2006.
10. I. Jeffery, D. Higgins, and A. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformatics*, vol. 7, p. 359, 2006.
11. A. Hossain, J. Beyene, A. Willan, and P. Hu, "A flexible approximate likelihood ratio test for detecting differential expression in microarray data," *Computational Statistics and Data Analysis*, vol. 53, pp. 3685–3695, 2009.
12. N. Eugene, C. Lee, and F. Famoye, "Beta-normal distribution and its applications," *Communication in Statistics-Theory and Methods*, vol. 31, pp. 497–512, 2002.
13. A. K. Gupta and S. Nadarajah, "On the moments of the beta-normal distribution," *Communication in Statistics- Theory and Methods*, vol. 33, pp. 1–13, 2004.
14. F. Famoye, C. Lee, and N. Eugene, "Beta-normal distribution: Bimodality properties and applications," *Journal of Modern Applied Statistical Methods*, vol. 3, pp. 85–103, 2004.
15. M. Razzaghi, "Beta-normal distribution in dose-response modeling and risk assessment for quantitative responses," *Environmental and Ecological Statistics*, vol. 16, pp. 25–36, 2009.
16. C. Broyden, "Quasi-newton methods and their application to function minimisation." *Mathematics of Computation*, vol. 21, pp. 368–381, 1967.
17. F. Cartieaux, M. Thibaud, L. Zimmerli, P. Lessard, C. Sarrobert, P. David, A. Gherbaud, C. Robaglia, S. Somerville, and L. Nussaume, "Transcriptome analysis of arabidopsis colonized by a plant-growth promoting rhizobacterium reveals a general effect on disease resistance," *The Plant Journal*, vol. 36, pp. 177– 190, 2003.
18. B. Craig, M. Black, and R. Doerge, "Gene expression data: The technology and statistical analysis," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 8, pp. 1–28, 2003.
19. W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 12, pp. 546–554, 2002.

20. Y. Zhao and W. Pan, "Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 19, pp. 1046–1054, 2003.

Appendix 1. Derivation of *BN*

Let $X_i = (X_{i1}, \ldots, X_{in_i}), i = 1, \ldots, N$ be the normalized log ratio of the expression levels in the two conditions for gene i. We assume X_{ij} is a random variable from a Normal distribution with mean μ_i and variance σ_i^2, $i = 1, \ldots, N$. Define a binary random variable I_i which equals 1 if the gene is differentially expressed under the two conditions; and takes the value 0 if the gene is not differentially expressed. In the case of $I_i = 1$, μ_i is assumed to be a random variable from the Beta-Normal distribution with mean 0 and variance $v\sigma_i^2$. In addition, we assume σ_i^2 is a random variable from the inverse Gamma distribution with rate parameter a and shape parameter b, so that the precision $\tau = 1/\sigma_i^2$ is a Gamma random variable with rate a and shape b. Thus we have

$$\begin{aligned}
&X_{i1}, \ldots, X_{in_i} | \mu_i, \tau_i \sim N(\mu_i, \sigma_i^2) = N(\mu_i, 1/\tau_i) \\
&\mu_i | \tau_i \sim \begin{cases} \delta(0), & I_i = 0 \\ f(u; 0, \sqrt{v}\sigma_i) = f(u; 0, \sqrt{v/\tau_i}), & I_i = 1 \end{cases} \\
&\tau_i \sim \Gamma(b, a),
\end{aligned}$$

the corresponding density functions are

$$\begin{aligned}
&f(X_i | \mu_i, \tau_i) = (2\pi)^{-\frac{n_i}{2}} \tau_i^{\frac{n_i}{2}} \exp\left\{ -\frac{\tau_i}{2} \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + n_i (\bar{X}_i - \mu_i)^2 \right] \right\}, \\
&f_{I=1}(\mu_i | \tau_i) = \frac{\sqrt{\tau_i}}{\sqrt{v} B(\alpha, \beta)} \left\{ \Phi\left(\frac{\mu_i \sqrt{\tau_i}}{\sqrt{v}} \right) \right\}^{\alpha - 1} \left\{ 1 - \Phi\left(\frac{\mu_i \sqrt{\tau_i}}{\sqrt{v}} \right) \right\}^{\beta - 1} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{\tau_i \mu_i^2}{2v} \right\}, \\
&f_{I=0}(\mu_i | \tau_i) = \delta(0), \\
&f(\tau_i) = \frac{a^b}{\Gamma(b)} \tau_i^{b-1} \exp\{-a\tau_i\}
\end{aligned}$$

Thus the log posterior odds for gene i can be found by calculating $f_{I_i=1}(X_i)$ and $f_{I_i=0}(X_i)$ respectively as follows, where $S_i^2 = 1/n_i \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$.

$$\begin{aligned}
f_{I_i=0}(X_i) &= \int_0^\infty \int_{-\infty}^\infty f(X_i | \mu_i, \tau_i) f_{I=0}(\mu_i | \tau_i) f(\tau_i) d\mu_i d\tau_i \\
&= \int_0^\infty f(X_i | 0, \tau_i) f(\tau_i) d\tau_i \\
&= \int_0^\infty (2\pi)^{-\frac{n_i}{2}} \tau_i^{\frac{n_i}{2}} \exp\left\{ -\frac{\tau_i}{2} \left[\sum_{j=1}^{n_i} (X_{ij})^2 \right] \right\} \frac{a^b}{\Gamma(b)} \tau_i^{b-1} \exp\{-a\tau_i\} d\tau_i \\
&= (2\pi)^{-\frac{n_i}{2}} \frac{a^b}{\Gamma(b)} \int_0^\infty \tau_i^{b + \frac{n_i}{2} - 1} \exp\left\{ -\frac{n_i \tau_i}{2} \left[S_i^2 + \bar{X}_i^2 \right] \right\} \exp\{-a\tau_i\} d\tau_i \\
&= (2\pi)^{-\frac{n_i}{2}} \frac{a^b}{\Gamma(b)} \int_0^\infty \tau_i^{b + \frac{n_i}{2} - 1} \exp\left\{ -\tau_i \left(a + \frac{n_i}{2} \left[S_i^2 + \bar{X}_i^2 \right] \right) \right\} d\tau_i \\
&= (2\pi)^{-\frac{n_i}{2}} \frac{a^b}{\Gamma(b)} \frac{\Gamma(b + \frac{n_i}{2})}{\left(a + \frac{n_i}{2} \left[S_i^2 + \bar{X}_i^2 \right] \right)^{b + \frac{n_i}{2}}},
\end{aligned}$$

and

$$f_{I_i=1}(X_i) = \int_0^\infty \int_{-\infty}^\infty f(X_i | \mu_i, \tau_i) f_{I=1}(\mu_i | \tau_i) f(\tau_i) d\mu_i d\tau_i$$

https://dx.doi.org/10.22606/jas.2016.14003

$$= \int_0^\infty \int_{-\infty}^\infty (2\pi)^{-\frac{n_i}{2}} \tau_i^{\frac{n_i}{2}} \exp\left\{-\frac{\tau_i}{2}\left[\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2 + n_i(\bar{X}_i-\mu_i)^2\right]\right\}$$
$$\times \frac{\sqrt{\tau_i}}{\sqrt{v}B(\alpha,\beta)}\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\tau_i\mu_i^2}{2v}\right\}$$
$$\times \frac{a^b}{\Gamma(b)}\tau_i^{b-1}\exp\{-a\tau_i\}$$
$$= (2\pi)^{-\frac{n_i+1}{2}}\frac{a^b}{\Gamma(b)}\frac{1}{\sqrt{v}B(\alpha,\beta)}\times$$
$$\int_0^\infty\int_{-\infty}^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}}\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\exp\left\{-\frac{\tau_i n_i}{2}S_i^2\right\}\exp\left\{-\frac{\tau_i n_i}{2}(\bar{X}_i-\mu_i)^2\right\}\exp\left\{-\frac{\tau_i\mu_i^2}{2v}\right\}\exp\{-a\tau_i\}d\mu_i d\tau_i$$
$$= (2\pi)^{-\frac{n_i+1}{2}}\frac{a^b}{\Gamma(b)}\frac{1}{\sqrt{v}B(\alpha,\beta)}\times$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}}\exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2\right]\right\}\int_{-\infty}^\infty\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\exp\left\{-\frac{\tau_i n_i}{2}(\bar{X}_i^2+\mu_i^2-2\bar{X}_i\mu_i)\right\}\exp\left\{-\frac{\tau_i\mu_i^2}{2v}\right\}d\mu_i d\tau_i$$
$$= (2\pi)^{-\frac{n_i+1}{2}}\frac{a^b}{\Gamma(b)}\frac{1}{\sqrt{v}B(\alpha,\beta)}\times$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}}\exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2\right]\right\}\int_{-\infty}^\infty\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\exp\left\{-\frac{\tau_i}{2v}\left[vn_i\bar{X}_i^2+(1+vn_i)\mu_i^2-2nv_i\bar{X}_i\mu_i\right]\right\}d\mu_i d\tau_i$$
$$= (2\pi)^{-\frac{n_i+1}{2}}\frac{a^b}{\Gamma(b)}\frac{1}{\sqrt{v}B(\alpha,\beta)}\times$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}}\exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2\right]\right\}\int_{-\infty}^\infty\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\exp\left\{-\frac{(1+vn_i)\tau_i}{2v}\left[\frac{vn_i\bar{X}_i^2}{1+vn_i}+\mu_i^2-2\frac{nv_i\bar{X}_i}{1+vn_i}\mu_i\right]\right\}d\mu_i d\tau_i$$

$$= (2\pi)^{-\frac{n_i+1}{2}}\frac{a^b}{\Gamma(b)}\frac{1}{\sqrt{v}B(\alpha,\beta)}\times$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}}\exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2\right]\right\}\int_{-\infty}^\infty\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\exp\left\{-\frac{(1+vn_i)\tau_i}{2v}\left[\frac{vn_i\bar{X}_i^2}{1+vn_i}-\frac{n^2v_i^2\bar{X}_i^2}{(1+vn_i)^2}\right]\right\}\exp\left\{-\frac{(1+vn_i)\tau_i}{2v}\left[\mu_i-\frac{nv_i\bar{X}_i}{1+vn_i}\right]^2\right\}d\mu_i d\tau_i$$
$$= (2\pi)^{-\frac{n_i+1}{2}}\frac{a^b}{\Gamma(b)}\frac{1}{\sqrt{v}B(\alpha,\beta)}\times$$
$$\int_0^\infty t\tau_i^{b+\frac{n_i}{2}-\frac{1}{2}}\exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2\right]\right\}\int_{-\infty}^\infty\left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\exp\left\{-\tau_i\frac{n_i\bar{X}_i^2}{2(1+vn_i)}\right\}\exp\left\{-\frac{(1+vn_i)\tau_i}{2v}\left[\mu_i-\frac{nv_i\bar{X}_i}{1+vn_i}\right]^2\right\}d\mu_i d\tau_i$$

$$= (2\pi)^{-\frac{n_i+1}{2}} \frac{a^b}{\Gamma(b)} \frac{1}{\sqrt{v}B(\alpha,\beta)} \times$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}} \exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2+\frac{n_i\bar{X}_i^2}{2(1+vn_i)}\right]\right\} \int_{-\infty}^\infty \left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1}$$
$$\left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1} \exp\left\{-\frac{(1+vn_i)\tau_i}{2v}\left[\mu_i-\frac{nv_i\bar{X}_i}{1+vn_i}\right]^2\right\} d\mu_i d\tau_i$$

Taking the ratio, the BN statistic is

$$BN_i = \log\left\{\frac{p}{1-p}\frac{\left[a+\frac{n_i}{2}(S_i^2+\bar{X}_i^2)\right]^{b+\frac{n_i}{2}}}{\sqrt{2v\pi}B(\alpha,\beta)\Gamma(b+\frac{n_i}{2})}\times\right.$$
$$\int_0^\infty \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}} \exp\left\{-\tau_i\left[a+\frac{n_i}{2}S_i^2+\frac{n_i\bar{X}_i^2}{2(1+vn_i)}\right]\right\} \int_{-\infty}^\infty \left\{\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\alpha-1} \left\{1-\Phi\left(\frac{\mu_i\sqrt{\tau_i}}{\sqrt{v}}\right)\right\}^{\beta-1}$$
$$\left.\exp\left\{-\frac{(1+vn_i)\tau_i}{2v}\left[\mu_i-\frac{nv_i\bar{X}_i}{1+vn_i}\right]^2\right\} d\mu_i d\tau_i\right\}$$

Appendix 2. The Relationship Between BN and B

Here we prove that the BN statistic based on the Beta-Normal distribution is a generalization of the B statistic of [8] and the latter is resulted when both parameters α and β are equal to 1. First we need to point out the relationship of parameters between the two settings. Under the [8] parameterization we have

$$\tau_i^* = \frac{n_i a^*}{2\sigma_i^2} \sim \Gamma(\nu, 1)$$
$$\mu_i|\tau_i^* \sim \begin{cases}\delta(0), & I_i = 0\\ N(0, \frac{cn_i a^*}{2\tau_i^*}), & I_i = 1\end{cases}$$
$$X_i \sim N(\mu_i, \frac{n_i a^*}{2\tau_i^*}),$$

while under our parameterization we have,

$$\tau_i = \frac{1}{\sigma_i^2} \sim \Gamma(b, a)$$
$$\mu_i|\tau_i \sim \begin{cases}\delta(0), & I_i = 0\\ f(\mu_i, 0, \sqrt{\frac{v}{\tau_i}}), & I_i = 1\end{cases}$$
$$X_i \sim N(\mu_i, \frac{1}{\tau_i}),$$

where f is the density of the Beta-Normal distribution, and so

$$\tau_i = \frac{2}{n_i a^*}\tau_i^* \sim \Gamma(\nu, \frac{n_i a^*}{2}),$$

with $a = n_i a^*/2$ and $b = \nu$. Furthermore, when $I_i = 0$, we consider the special case of Beta-Normal distribution as $\alpha = \beta = 1$, then we get $v = c$. We have

$$f_{I_i=0}(X_i)$$
$$= (2\pi)^{-n_i/2} \frac{(\frac{n_i a^*}{2})^\nu}{\Gamma(\nu)} \frac{\Gamma(\nu+\frac{n_i}{2})}{[\frac{n_i a^*}{2}+\frac{n_i}{2}(S_i^2+\bar{X}_i^2)]^{\nu+\frac{n_i}{2}}}$$

https://dx.doi.org/10.22606/jas.2016.14003

$$= \frac{\Gamma(\nu + \frac{n_i}{2})}{\Gamma(\nu)}(2\pi)^{-\frac{n_i}{2}}(\frac{n_i a^*}{2})^{-\frac{n_i}{2}}[1 + \frac{1}{a^*}(S_i^2 + \bar{X}_i^2)]^{-(\nu+\frac{n_i}{2})},$$

and for $I_i = 1$ with $\alpha = \beta = 1$, considering the Beta function $B(1,1) = 1$,

$$\begin{aligned}
& f_{I_i=1}(X_i) \\
&= (2\pi)^{-\frac{n_i+1}{2}} \frac{a^b}{\Gamma(b)} \frac{1}{\sqrt{v}} \times \\
& \int \tau_i^{b+\frac{n_i}{2}-\frac{1}{2}} \exp\left\{-\tau_i \left[a + \frac{n_i}{2} S_i^2 + \frac{n_i \bar{X}_i^2}{2(1+vn_i)}\right]\right\} \sqrt{2\pi} \sqrt{\frac{v}{1+vn_i}} \frac{1}{\sqrt{\tau_i}} d\tau_i \\
&= (2\pi)^{-\frac{n_i}{2}} \frac{a^b}{\Gamma(b)} \frac{1}{\sqrt{1+vn_i}} \frac{\Gamma(b+\frac{n_i}{2})}{\left[a + \frac{n_i}{2} S_i^2 + \frac{n_i \bar{X}_i^2}{2(1+vn_i)}\right]^{b+\frac{n_i}{2}}} \\
&= \frac{\Gamma(\nu + \frac{n_i}{2})}{\Gamma(\nu)} (2\pi)^{-\frac{n_i}{2}} \left(\frac{n_i a^*}{2}\right)^{-\frac{n_i}{2}} (1+cn_i)^{-1/2} \left[1 + \frac{1}{a^*}\left(S_i^2 + \frac{\bar{X}_i^2}{1+cn_i}\right)\right]^{-(\nu+\frac{n_i}{2})},
\end{aligned}$$

Therefore, we can conclude when $\alpha = \beta = 1$, B in [8] is resulted from the proposed BN statistic.

