A MIXTURE MODEL APPROACH FOR GENE SELECTION USING JOHNSON'S SYSTEM AND BAYES FORMULA

FLORENCE GEORGE AND KANDETHODY M. RAMACHANDRAN

Department of Statistics, Florida International University, 11200 S.W.8th Street, Miami, FL-33199, USA.

Department of Mathematics and Statistics, University of South Florida(USF), 4202 E. Fowler Ave, Tampa, FL-33620, USA.

ABSTRACT. Microarrays have become increasingly common in biological and medical research. They enable the simultaneous study of thousands of genes and provide gene expression information on a whole genome level. A major goal of microarray experiments is to determine which genes are differentially expressed between samples. A mixed model approach using the Johnson's system of distributions and Baye's formula is proposed in this paper for the selection of differentially expressed genes. In this approach, no specific parametric distribution is assumed for the gene expression levels. The simulation results show that the proposed approach has a higher power over the other commonly used Bayesian methods such as EBarrays and EBAM.

Key Words Microarrays,Gene selection, Differentially expressed genes, Johnson's Distribution.

1. INTRODUCTION

While simultaneous measurement of thousands of gene expression levels provides a potential source of profound knowledge, success of the microarray technology depends heavily on statistical analysis. Careful statistical thinking and analysis are required to find the underlying structure in the data. The unprecedented amounts of data produced by microarrays raise new challenges for statisticians to be able to perform inference on a scale never before conducted. Recently, statisticians and researchers in bioinformatics have focused much attention on the development of statistical methods to identify differentially expressed genes, with special emphasis on those methods that identify genes that are differentially expressed between two conditions. This work focusses on the development of a statistical method that is suitable for differential gene selection using Johnson system of distributions and Bayes's formula.

In contrast to methods that apply classical statistical inferences separately for different genes, there is a kind of information sharing among genes in mixture model

Received xxxx, 200x

analysis using Baye's formula. This can be beneficial because the data from other genes provide some information about the typical variability in the system[9],[12]. In this paper we will discuss two well known mixture model approaches using Baye's formula and introduce application of Johnson's system of distribution in the mixture model setup for the selection of differentially expressed genes.

2. DATA

Ovarian cancer is the fifth leading cause of cancer death among women in the United States and Western Europe, and has the highest mortality rate of all gynaecologic cancers. Currently, the standard treatment protocol used in the initial management of advanced-stage ovarian cancer is primary cytoreductive surgery followed by primary platinum-based chemotherapy. However, approximately 30% of patients with advanced stage disease do not demonstrate a complete response to primary platinum-based therapy. Identifying genes which are expressed significantly different in the two groups, could provide some insight for the precise diagnosis of response to the treatment and help the medical specialists to choose an alternate therapy when needed. The ovarian cancer tissue samples involved in this study are collected from the tumor banks at the H.Lee Moffitt Cancer Centre & Research Institute and Duke University Medical center. Affymetrix U133A Gene Chip arrays were used to measure expression of 22,283 genes in advanced stage serous ovarian cancers from 55 patients who underwent primary surgery followed by platinum-based chemotherapy. Expression values are calculated using the robust multi-array (RMA) algorithm[5] implemented in the Bioconductor ($http:\backslash\backslash www.bioconductor.org$) extensions to the R statistical programming environment[4]. Gene expressions were compared between patients who demonstrated a complete response to platinum-based therapy and those who did not to identify differentially expressed genes.

2.1. **Data Simulation.**

The ovarian cancer data is used as the target model for simulation. We use the approach discussed in [6] for data simulation. Given the parameters, gene expression are generated randomly from a gamma distribution. But for each gene, the parameters are generated randomly. The means of gene expressions are generated from a normal distribution $N(\mu, \sigma)$ and standard deviations from a gamma distribution $Gamma(\alpha, \beta)$. The hyper parameters μ, σ, α and β are chosen to fit the ovarian cancer data. These values are $\mu = 6.7, \sigma = 1.68, \alpha = 8.76$ and $\beta = 9.386$. The parameters $(shape = \alpha_i, rate = \beta_i)$ for gene i are calculated from the generated mean μ_i and standard deviation σ_i using the relations $\alpha_i = \mu_i^2/\sigma_i^2$ and $\beta_i = \mu_i/\sigma_i^2$ respectively. Gene expressions for 1,000 genes were simulated. The number of replications are selected as unequal as in studies like ovarian cancer data, where treatment response of patients under similar conditions are of interest, it is more likely to get samples of different sizes. Data sets are simulated (1)with 15 replication in the first group, 10 replications in the second group and (2)with 33 replication in the first group, 22 replications in the second group. We choose 5% of the genes to be differentially expressed. For differentially expressed genes the parameters are chosen to be different in the two groups.

3. JOHNSON'S SYSTEM OF DISTRIBUTIONS

In 1949, N.L. Johnson derived a system of curves[2] that had the flexibility of covering a wide variety of shapes and had practical advantage of being able to transform to the normal distribution. This system contains three families of distributions; Bounded form S_B, Unbounded form S_U and Log Normal form S_L. The Johnson system is able to closely approximate many of the standard continuous distributions through one of three functional forms and is thus highly flexible. The three systems of Johnson's family are generated by the following transformations of a continuous random variable x to the standard normal variable z.

(a)*Log Normal form S_L.* The transformation here is,

$$z = \gamma + \eta ln\left(\frac{x-\epsilon}{\lambda}\right), \tag{3.1}$$

(b)*Bounded form S_B.* It is the set of distributions that have a fixed boundary on either the upper or lower tail, or both. The transformation is,

$$z = \gamma + \eta ln\left(\frac{x-\epsilon}{\epsilon+\lambda-x}\right), \epsilon < x < \epsilon + \lambda \tag{3.2}$$

(c)*Unbounded form S_U.* It is the set of distributions that go to infinity in both the upper and lower tail. The transformation is,

$$z = \gamma + \eta sinh^{-1}\left(\frac{X-\epsilon}{\lambda}\right) \tag{3.3}$$

where the parameters γ, η, λ, ϵ are to be estimated using data values. See [2] for further description of these distributions. The chosen functions are such that in a plot of the third and fourth standardized moments, β_1(measure of skewness) and β_2(measure of kurtosis), the S_L distribution form a curve which divides the (β_1, β_2) plane into two regions. The S_B distribution lie in one of the regions and the S_U lie in the other. When using the Johnson system, the first step is to determine which of the three families should be used. The usual procedure is to compute the sample estimates of the standardized moments and choose the distribution according to which of the two regions the computed point falls into[2]. The selection of Johnson's system and estimation of parameters by using sample quantiles [13] is introduced by Wheeler. Slifker and Sapiro introduced another selection rule which is a function of four percentiles for selecting one of the three families and to give estimates of the parameters[10] as explained below.

Choose any fixed value ζ between 0 and 1. Then the four points $\pm\zeta$ and $\pm 3\zeta$ determine three intervals of equal length. Determine the standard normal percentile P_ζ corresponding to $z = 3\zeta, \zeta, -\zeta, -3\zeta$ respectively. For example if $\zeta = 0.5$ then using standard normal tables,$P_{-1.5} = 0.0668 * 100 = 6.68$,$P_{-0.5} = 0.3085 * 100 = 30.85$, $P_{0.5} = 0.6915 * 100 = 69.15$, $P_{1.5} = 0.9332 * 100 = 93.32$. Let$x_{3\zeta}, x_\zeta, x_{-\zeta}, x_{3\zeta}$ be the percentiles of data values corresponding to the four selected percentiles of the Normal distribution. The type of distribution chosen is based on the value of the discriminant d calculated a s follows.

$$d = \frac{mn}{p^2} \tag{3.4}$$

where $p = x_\zeta - x_{-\zeta}$, $m = x_{3\zeta} - x_\zeta$, $n = x_{-\zeta} - x_{-3\zeta}$. If the calculated discriminant d is greater than 1.001, then an unbounded distribution is chosen. If the value is less than 0.999, then a bounded distribution is chosen. A discriminant equal to or between the two values results in a Log Normal fit. The fit parameters for the transformation are calculated by solving the transformation equation for the chosen distribution type at the four selected percentiles.

The flexibility provided by the choice of form and fitting parameters allows for great flexibility in adjusting the curve to fit the data. The fact that the Johnson system involves a transformation of the raw variable to a Normal variable allows estimates of the percentiles of the fitted distribution to be calculated from the Normal distribution percentiles.

4. **EBARRAYS AND EBAM**

Two well known mixture model approaches for the selection of differentially expressed genes are EBarrays and EBAM. Parametric Empirical Bayes(EBarrays) approach developed by Kendziorski *etal* computes the posterior probability under one of the two proposed hierarchical model assumption of the expression levels, one based on the assumption of Gamma distributed measurements(EBarrays-GG) and the other based on log-normally distributed measurements(EBarrays-LNN)[6].

EBAM assumes that there are two classes of genes namely ”Different” and ”Not Different” meaning that the gene is either differently or not differently expressed in two different groups under consideration. This will give two possible probability distributions for any summarizing value of a gene which is able to measure the difference in expression levels of the genes in the two groups. EBAM developed by Efron [1] use the t-value defined by the the following equation to summarize the information about any gene.

$$t_i = \frac{\bar{x}_{2i} - \bar{x}_{1i}}{\sqrt{\frac{r_1 S_{1i}^2 + r_2 S_{2i}^2}{r_1 + r_2 - 2}\left(\frac{1}{r_1} + \frac{1}{r_2}\right)}} \tag{4.1}$$

Let $f_0(y)$ be the density of the summary value y for equally expressed genes and $f_1(y)$ be the density of y for differentially expressed genes. Let the prior probabilities for the two classes be p_0 and $p_1 = 1 - p_0$ with the corresponding densities $f_0(y)$ and $f_1(y)$ respectively. Hence we have the marginal density f(y) for y, which is a mixture density of the two populations as,

$$f(y) = p_0 f_0(y) + p_1 f_1(y) \tag{4.2}$$

EBAM uses Bayes theorem to obtain posteriori probabilities of any gene to be differentially expressed.

$$\begin{aligned} p_0(y) &= prob(Equally \;\; expressed|Y = y) \\ &= \frac{prob(Equally \;\; expressed \;\; and \;\; Y = y)}{p(Y = y)} \\ &= \frac{prob(Y = y|Equally \;\; expressed) \times prob(Equally \;\; expressed)}{p(Y = y)} \\ &= \frac{p_0 f_0(y)}{f(y)} \qquad (4.3) \end{aligned}$$

and

$$p_1(y) = prob(Differentially \;\; expressed|Y = y) = 1 - \frac{p_0 f_0(y)}{f(y)} \tag{4.4}$$

To estimate the posterior probabilities we need $f_0(y)$, $f(y)$ and p_0. Efron estimated $f_0(y)$ as the t-density with appropriate degrees of freedom. The mixture density $f(y)$ is estimated by fitting a smooth curve $\hat{f}(y)$ to the Y histogram.

5. GENE SELECTION USING JOHNSON DISTRIBUTION AND BAYE'S FORMULA

Here we modify EBAM using Johnson's system of distributions. The Baye's formula is used to incorporate the overall information about the analytical characteristics of genes to identify differentially expressed genes. Instead of using t-value, we use m-value defined by the following equation to summarize the information about any gene. we define m-value by

$$m_j = \frac{(\bar{x}_{j2} - \bar{x}_{j1})}{s_j + a_0} \tag{5.1}$$

where

$$s_j = \sqrt{\frac{var(x_{j1})}{n_1} + \frac{var(x_{j2})}{n_2}} \tag{5.2}$$

and a_0 is a shrinkage parameter which depends on the s_j values. The value m_j is the $m - value$ for gene j. The constant a_0 in the denominator of Equation 5.1 can lead to the reduction of the overall variance of the m_j, giving the tests more power on average. This has the added effect of dampening large values of t-statistics that arise from small variance of genes. We have taken a_0 as the median of the s_j values.

Then $f_0(m)$ is the distribution of the m-values when the genes are equally expressed. The balanced permutation technique and Johnson's system of distributions are used to estimate $f_0(m)$. More specifically, we will create artificial groups by taking permutations of the microarray samples and randomly assign one of the two labels to each of these groups. Then calculate the m-values for this artificial groups. We did 50 permutations and use the m-values from these 50 permutations to estimate $f_0(m)$. The estimated $f_0(m)$ is an unbounded Johnson distribution with parameters $\hat{\gamma} = 0.1059483$, $\hat{\delta} = 2.690686$, $\hat{\xi} = 0.05775453$ and $\hat{\lambda} = 1.301382$. We can estimate $f(m)$ empirically using Johnson distribution, as any continuous distribution can be approximated by a Johnson distribution. The number of calculated statistics is the same as the number of genes, large enough to estimate the empirical distribution. The marginal distribution $f(m)$ is estimated as a bounded Johnson distribution with parameters $\hat{\gamma}$= -1.127296, $\hat{\delta}$ = 1.41565 , $\hat{\xi}$ = -3.071095 and $\hat{\lambda}$ = 4.680507. These parameters of the Johnson's distribution are estimated using Quantiles method [13].

The value of p_0 chosen in such a way that all posterior probabilities are positive [1]. We make use of the estimated value of p_0 used in EBAM [1] under the same criterion. Now we are able to calculate the posterior probability using the Equation 4.4. The genes with posterior probabilities greater than 0.8 are chosen as the differentially expressed genes. The Figure 1 shows posterior probabilities against m-values. The gray spots corresponds to the genes with posterior probability > 0.8.

The Kolmogrov-Smirnov goodness of fit test is done to see how fit the estimated Johnson distributions are for the corresponding observed values of the statistics. The p-value for the null distribution is 0.4569 and the p-value for the marginal distribution is 0.15693.

Both EBArrays and EBAM share information among genes. One drawback of EBArrays is the assumption of a parametric model for gene expressions and hence a

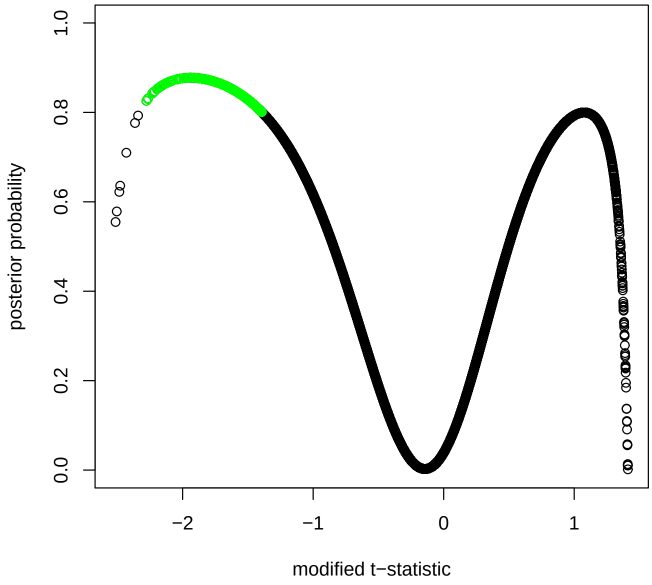

FIGURE 1. Posterior probability of genes; Genes representing green(or gray) points have posterior probability > 0.8

good chance of violation of assumptions. EBArrays assumes that the gene expressions are generated from a Gamma distribution or from a Log-Normal distribution. EBAM assumes a t-distribution as the distribution of t-values for equally expressed genes, which in turn requires, the assumptions of t-distribution to hold. In the method we proposed, we improve EBAM using m-values and Johnson's system of distribution. The distribution of the m values is estimated using Johnson's system of distributions. The advantage of this proposed method is that while sharing information across all of the genes, there is no parametric assumption on the gene expression data. Here we make use of the fact that any continuous distribution is a special case of Johnson system of distribution [2]. The simulation shows that the proposed method is better than EBArrays and EBAM(Figures 2 to 7).

6. **RESULTS**

The proposed methods using Johnson's distribution are compared with EBarrays, EBAM and SAM. Significance Analysis of Microarrays"(SAM) [11] is the most popular classical method employed for microarray data analysis. The number of genes selected as differently expressed, out of 22,283 genes, by these methods are listed in Table 1.

Method	**No.of genes**
Bayes -Johnson	543
EBArrays - GG	191
EBArrays - LNN	177
EBAM	223
SAM	2166

Table 1. No. of genes selected by different methods

Because no truth about differentially expressed genes could be obtained on ovarian cancer data, it is not possible to compare results obtained for the real data. In order to assess the effectiveness of the proposed methodology and to obtain a quantitative evaluation of gene selection methods, the simulated data explained in Section 2.1 is used. A comparison of methods discussed in this paper is presented in Figures 2 to 7. The number of truly differentially expressed genes are plotted against the number of genes selected in Figures 2 to 4. We can observe that for any value of number of genes selected(x-coordinate), the proposed method using Johnson system of distributions gives the more number of true positives than other methods. The proposed method using the Johnson system of distributions seem to be superior to other methods as the number of truly differentially expressed genes identified is more than that identified by other methods, for any fixed number of genes selected. After some saturation point where all the truly differentially expressed genes are identified these curves will converge.

The results of the methods discussed here are also displayed by Receiver Operating Characteristic (ROC) curves in Figures 5 to 7 using the simulated data. The ROC curve displays the false positive rate (rate of non-Differentially Expressed Genes(non-DEGs) included) versus the false negative rate (rate of DEGs not included). The false positive rate is the proportion of number of Equally expressed genes that were erroneously reported as Differentially Expressed. Hence $False\ \ positive\ \ rate = \frac{Number\ \ of\ \ false\ \ positives}{Number\ \ of\ \ Equally\ \ Expressed\ \ genes}$. This is the same as the probability of Type I error denoted by α. The false negative rate is the proportion of Differentially Expressed genes that were erroneously reported as Equally Expressed. More specifically, $False\ \ Negative\ \ Rate = \frac{Number\ \ of\ false\ \ negatives}{Number\ \ of\ \ Differentially\ \ expressed\ \ genes}$. This is the same as the probability of type II error. It is equal to 1 minus power of the test. A method whose ROC curve lies below another one is preferred [7], as the curve represents the

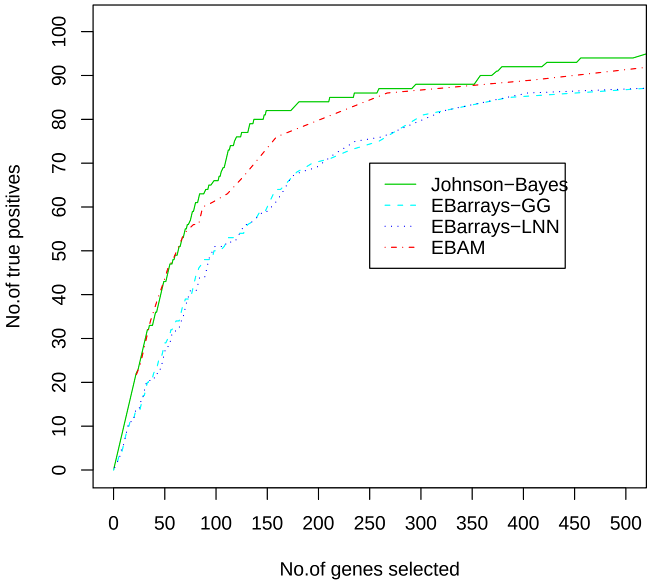

Figure 2. Mixture model approaches - No. of genes selected vs No. of True Positives; Number of genes -2000; Sample size - 15 vs 10

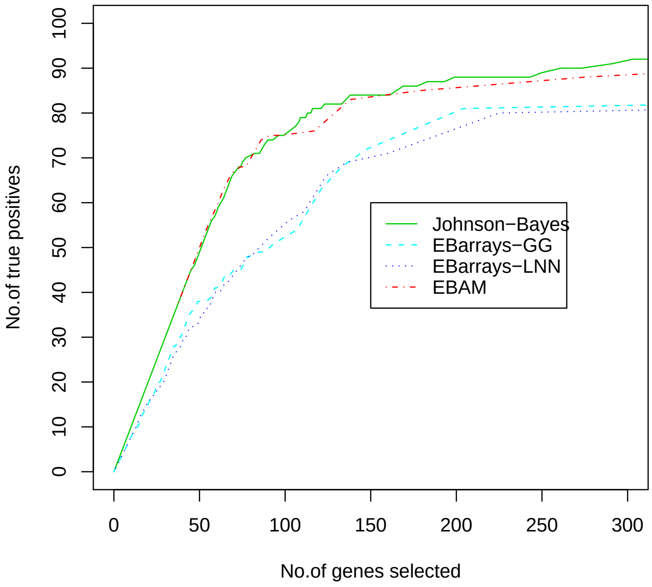

Figure 3. Mixture model approaches - No. of genes selected vs No. of True Positives; Number of genes -2000; Sample size - 20 vs 20

Type I and Type II errors. A method which has a better ROC curve, in this sense, will produce top lists with more differentially expressed genes(DEGs), fewer non-DEGs and consequently, will leave out fewer DEGs. It can be observed from the Figures 5 to 7 that the proposed approach using Johnson's system of distributions is better than the other methods discussed here.

7. SUMMARY

Here we used Baye's formula and Johnson system of distribution together with m values defined by Equation 5.1 to identify differentially expressed genes. Johnson system is used to approximate the probability distribution of the summary measure($m-value$). Then Bayes formula is used to revise the probability of each gene to be differentially expressed. As in previous chapter the method is applied for the gene selection of ovarian cancer data. A comparison study of the method is done with the existing methods. For comparison purposes we have used data simulated using the information from the real data. We have identified the distribution that characterizes the real data and obtained the maximum likelihood estimates using this data. Then these estimates are utilized to numerically simulate the information. The simulation study shows that the proposed method using Johnson's system of distribution is better than the popular mixture model methods EBAM and EBArrays .

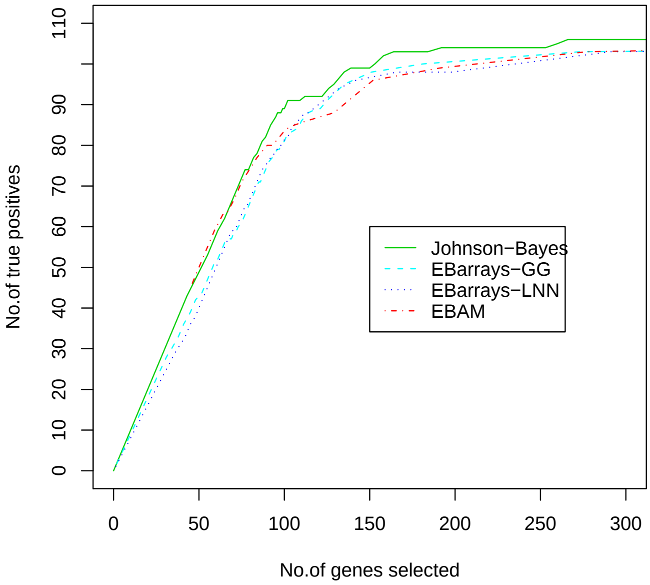

FIGURE 4. Mixture model approaches - No. of genes selected vs No. of True Positives; Number of genes -2000; Sample size - 33 vs 22

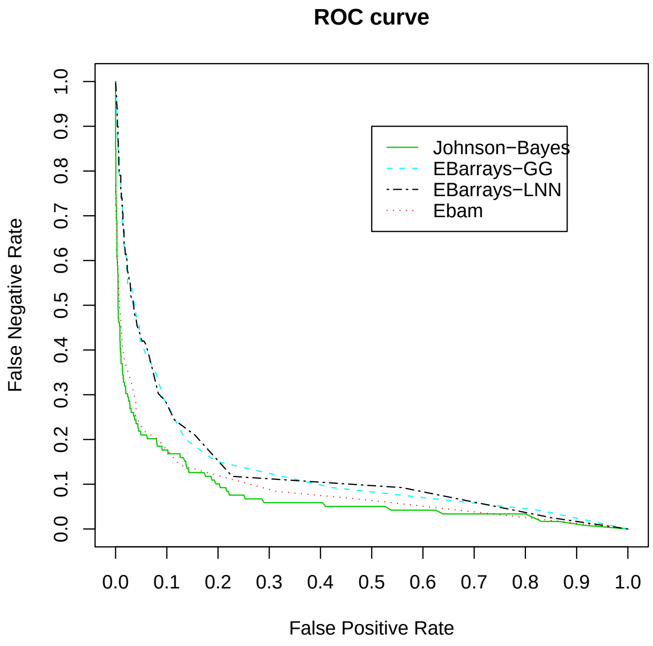

FIGURE 5. ROC curve -Mixture model approaches; Number of genes -2000; Sample size - 15 vs 10

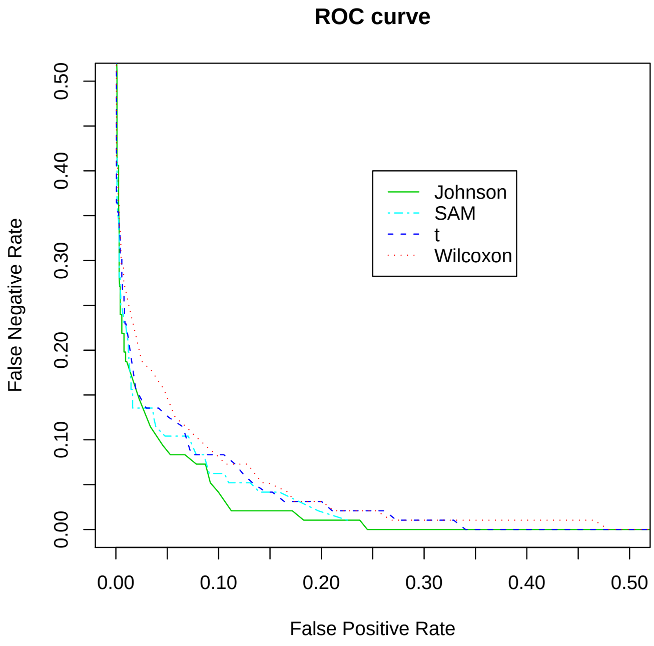

FIGURE 6. ROC curve -Mixture model approaches; Number of genes -2000; Sample size - 20 vs 20

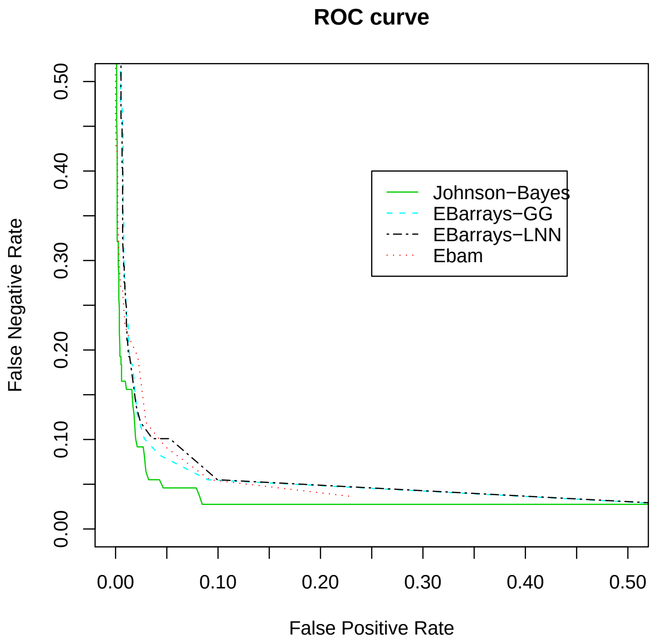

Figure 7. ROC curve -Mixture model approaches; Number of genes -2000; Sample size - 33 vs 22

REFERENCES

[1] Efron B. Robbins, Empirical Bayes and Microarrays. The annals of Statistics, 31:366-378, 2003.

[2] Johnson, N.L., Systems of frequency curves generated by methods of translation. Biometrika 36, 149-176,1949.

[3] Goulb T.R., Slonim, D.K. Tamayo P. *et.al.*., Molecular classification of cancer class discovery and class prediction by gene expression monitoring. Science 286, 531-537,1999.

[4] Ihaka R, Gentleman R., R:A language for data analysis and graphics", J.Comput.Graph.,5:299-314, 1996.

[5] Irizarry RA, Hobbs B, Collin F., Beazer-Barely YD, Antonellis KJ, Scherf U et al., Exploration, Normlization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics,4(2): 249-264, 2003.

[6] Kendziorski, C.M., M.A. Newton, H. Lan, and M.N. Gould, On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine; 22: 3899-3914, 2003.

[7] Lonnstedt I, Speed TP, Replicated Microarray data. Stat Sinica, 12:31-46, 2002.

[8] Mei-Ling Ting Lee, Analysis of gene expression data, Kluwer Academic publishers, 2004.

[9] Parmigiani G., E.S. Garrett, R.A.Irizarry, S.L.Zeger, The Analysis of Gene Expression Data. Springer, Newyork, 2003

[10] Slifker J, Shapiro S, The Johnson System : selection and parameter estimation. Technometrics; 22:239-247, 1980.

[11] Tusher V., Tibshirani R., Chu C., Significance Analysis of microarrays applied to transcriptional response to ionizing radiation., Proeedings of the National Academy of Sciences;98: 5116-5121, 2001.

[12] Vladimir P. Savchuk, and Chris P. Tsokos, Bayesian Statistical Methods with Applications to Reliability, World Federation Publishers, Inc., 1996.

[13] Wheeler R, Quantile Estimators of Johnson curve Parameters. Biometrika vol.67 No.3 pp 725-728, 1980.