*Journal of Statistical Research* *ISSN 0256 - 422 X*
*2009, Vol. 43, No. 1, pp. 117-125*
*Bangladesh*

# GENE SELECTION WITH JOHNSON'S DISTRIBUTION

Florence George
*Florida International University*
*Department of Statistics, Miami, Florida, USA*
*Email: fgeorge@fiu.edu*

K.M. Ramachandran
*University of South Florida*
*Department of Mathematics and Statistics, Tampa, Florida, USA*
*Email: ram@cas.usf.edu*

Li Lihua
*Institute for Biomedical Engineering and Instrumentation,*
*Hangzhou Dianzi University, China*
*Email: lilh@hdu.edu.cn*

SUMMARY

Microarrays have become increasingly common in biological and medical research. A major goal of microarray experiments is to determine which genes are differentially expressed between samples. A new approach using the Johnson's system of distributions is proposed in this paper to make simultaneous inference concerning which genes are differentially expressed, in which no specific parametric distribution is assumed for the gene expression levels. The simulation study shows that the new approach has better results compared to many existing methods.

*Keywords and phrases:* gene selection, Differentially expressed genes, Johnson's Distribution, microarrays

## 1 Introduction

Microarrays are devices for measuring gene *expression levels*, that is, how active a particular gene is in the working of a given cell. Microarrays provide the expression levels of thousands of genes simultaneously. Microarray technology can provide important insights about the underlying genetic causes of many important biological questions. One of the major goals of microarray data analysis is the identification of genes that are differentially expressed across two or more samples under different experimental conditions. Genes that show differential expression under different experimental conditions will allow for the identification of biomarkers for disease class predictions as well as the ability to fine-scale predictions of

drug responses. A large number of methods have been developed for the selection of differentially expressed genes. Many methods that have been proposed to assess differential analysis are based on using the two-sample t-test or a minor variation of the t-statistic, but they differ in how to associate a statistical significance level ($p-value$) to the corresponding summary statistic (13). The $p-value$ is calculated based on the distribution of the test statistic under the null hypothesis. Differences in how a significance level is assigned could lead to possibly large differences in the numbers of genes detected and the number of false positives and false negatives.

A straight forward method for the selection of differentially expressed genes is the traditional $t$-test (2). Under the normality assumption of the expression levels, the $t$-statistic follows a $t$-distribution in a standard $t$-test. Tusher *et al.* (15) proposed the Significance Analysis of Microarrays(SAM) version of the $t$-tests. In SAM (15), instead of regular $t$-statistic, *modified-t* is used where the denominator is variance plus a fudge factor (see the original paper for details). There is no parametric assumption like normality on expression levels. Essentially SAM identifies genes with statistically significant changes in expressions by assimilating information from a set of gene-specific modified-$t$ statistic and deciding whether this statistic is significant based on permutations of the available experimental data. The size of the data set should be large enough to allow for a sufficient number of distinct permutations to be obtained. The Wilcoxon rank sum test has also been used as alternative to the $t$-test in the two-sample comparison of microarray data (11). As it is non-parametric, there are no parametric assumptions but it assumes the two samples are derived from identical distributions, with the only difference being their location parameters. Parametric Empirical Bayes(EBarrays) approach developed by Kendziorski *et al* 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) (9), (12). In addition to the assumption of specific parametric model, a constant coefficient of variation of expression levels is also assumed in EBarrays.

This paper discusses a new approach to identify differentially expressed genes using the Johnson's system of distributions and compare the results with some commonly used methods. We use Johnson's distribution to estimate the null distribution without any parametric assumption of gene expression levels.

# 2 Method

## 2.1 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 differently 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 (6) implemented in the Bioconductor (`http:\\www.bioconductor.org`) extensions to the R statistical programming environment (5). 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.

## 2.2 Johnson's System of Distributions

In 1949, Johnson derived a system of curves (7),(8) that has the flexibility of covering a wide variety of shapes. The Johnson's system has the practical and theoretical advantages of being able to transform to the normal distribution (7). The Johnson system is able to closely approximate many of the standard continuous distributions through one of the three functional forms and is thus highly flexible. The Johnson system provides one distribution corresponding to each pair of mathematically possible values of skewness and kurtosis.

The significant flexibility of Johnson system of distributions is very useful in characterizing the complicated data set like microarray data. In all parametric approaches of identifying significant genes, there is a distributional assumption for gene expression like Normal, log-normal, Gamma etc. The advantage of using Johnson system of distributions is that many, if not all, of the commonly used continuous distributions such as Normal, log-normal, Gamma, Beta, Exponential are special cases of Johnson system (4). The Johnson system give a variety of shapes of curve as wide as that provided by the systems of frequency curves in general use (7). Any continuous distribution with finite moments can be approximated by a member of the Johnson's system (7). This motivated us using Johnson system for the analysis of microarray data.

Given a continuous random variable $X$ whose distribution is unknown and is to be approximated, Johnson proposed three normalizing transformations having the general form

$$Z = \gamma + \delta f\big((X - \xi)/\lambda\big),$$

where $f(.)$ denotes the transformation function, $Z$ is a standard normal random variable, $\gamma$ and $\delta$ are shape parameters, $\lambda$ is a scale parameter and $\xi$ is a location parameter. Without loss of generality, it is assumed that $\delta > 0$ and $\lambda > 0$. The first transformation proposed by

Johnson defines the lognormal system of distributions denoted by $S_L$:

$$\begin{aligned} Z &= \gamma + \delta \ln\big((X-\xi)/\lambda\big),\ X > \xi \\ &= \gamma^* + \delta \ln(X-\xi),\ X > \xi. \end{aligned}$$

$S_L$ curves cover the lognormal family.

The bounded system of distributions $S_B$ is defined by

$$Z = \gamma + \delta \ln\big((X-\xi)/(\xi+\lambda-X)\big),\ \xi < X < \xi + \lambda.$$

$S_B$ curves cover bounded distributions. The distributions can be bounded on either lower end, the upper end, or both. This family covers gamma distributions, beta distributions and many others.

The $S_U$ curves are unbounded and cover the $t$ and normal distributions, among others. The unbounded system of distributions $S_U$ is defined by

$$\begin{aligned} Z &= \gamma + \delta \ln\left[\left(\frac{X-\xi}{\lambda}\right) + \left\{\left(\frac{X-\xi}{\lambda}\right)^2 + 1\right\}^{1/2}\right],\ -\infty < X < \infty \\ &= \gamma + \delta \sinh^{-1}\left(\frac{X-\xi}{\lambda}\right), \end{aligned}$$

where $\gamma, \eta, \epsilon, \lambda$ are the parameters to be estimated using data values. In a plot of the third and fourth standardized moments, $\beta_1$(measure of skewness) and $\beta_2$(measure of kurtosis), the $S_L$ distribution form a curve which divides the $(\beta_1, \beta_2)$ 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 (7). A method for the selection of Johnson's system and estimation of parameters by using sample quantiles (16) is introduced by Wheeler. Slifker and Shapiro 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 (14).

## 2.3 Johnson's Distributions for Gene Selection

We are interested in determining which genes show a statistically significant difference in gene expression between two conditions. Consider the situation where there are $n_1$ replicate samples for condition-1 and $n_2$ replications of condition-2. Summarize the information on gene $j$ to the $m$-value defined by

$$m_j = \{(\bar{x}_{j2} - \bar{x}_{j1})/(s_j + a_0)\}, \tag{2.1}$$

where $s_j = \sqrt{\{var(x_{j1})/n_1\} + \{var(x_{j2})/n_2\}}$ and $a_0$ is a shrinkage parameter which depends on the $s_j$ values. The constant $a_0$ in the denominator of Equation 2.1 can lead to the

reduction of the overall variance of $m_j$, giving the tests more power. This has the added effect of dampening large values of the statistic that may arise from small variance of genes. We have taken $a_0$ as the median of the $s_j$ values. The idea of modifying estimators of variance has been presented by others in similar contexts. The SAM $t$-test (15) adds a small constant to the gene-specific variance estimate in order to stabilize the small variances. In SAM, the fudge factor $s_0$ is chosen as the value which minimizes the coefficient of variation of the SAM statistic $d_i$. The regularized $t$-test proposed by Baldi and Long (1) replaces the usual variance estimate with a Bayesian estimator based on hierarchical prior distribution.

For each gene we calculate the $m$-value using Equation 2.1. To make a decision about the significance of the summarized value of any gene, we need to know the null distribution $f_0$ of $m$, when genes are equally expressed. For this purpose we make use of permutation technique and Johnson's distribution as follows. Randomly select $n_1$ units from the combined pool of $n = n_1 + n_2$ sample units and label them as group 1. The remaining $n_2$ units will be labeled as group 2. Now calculate the $m$-value using Equation 2.1. This procedure is repeated a sufficient number of times. For each permutation, compute the $m$-values defined in Equation 2.1 and this can be considered as realizations of the $m$-values when the genes are equally expressed. The null hypothesis $H_0$ for a microarray study states that none of the genes is differentially expressed. Under this $H_0$, it is plausible to assume that the $m$-values derived from permutations of group labels are drawn independently from a common distribution with some probability density function, say, $f_0$. Johnson in his paper claims that, the Johnson's system of curves include most, if not all, continuous distributions that encountered in any collected data (7). Therefore to estimate the distribution $f_0$ we can assume that $f_0$ is a member of Johnson's system. Wheeler's quantile method is used for fitting the Johnson system. When we apply this method to the ovarian cancer data, an unbounded Johnson's distribution is fitted to form the common null distribution of the $m$-statistic, with parameters $\hat{\gamma} = 0.1059$, $\hat{\delta} = 2.6907$, $\hat{\xi} = 0.05775$ and $\hat{\lambda} = 1.3014$. Now the p-values of the calculated $m$-values can be obtained using this estimated distribution $f_0$. We have to fix a cut-off point for the $p$-values to select the differentially expressed genes. There were 442 genes in the ovarian cancer data with $p - value < 0.01$.

## 2.4 Data Simulation

Simulation studies were done in order to assess the effectiveness of the proposed methodology and to obtain a quantitative evaluation of gene selection methods.

The ovarian cancer data is used as the target model for simulation. We use the approach discussed in (9) 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 hyperparameters $\mu, \sigma, \alpha$ and $\beta$ are chosen to fit the ovarian cancer data. These values are $\mu = 6.7, \sigma = 1.68, \alpha = 8.76$ and $\beta = 9.386$. Gene expressions for 2,000 genes were simulated. Five data sets are simulated with (1)15 replication in the first group, 10 replications in the

second group; (2)33 replication in the first group, 22 replications in the second group; (3)20 replications in each group; (4)30 replications in each group and (5)40 replications in each group. We choose 5% of the genes to be differentially expressed.

### 2.5 Results and Discussion

The merit of the method depends on the ability to successfully identify differentially expressed genes while avoiding to classify unchanged genes as differentially expressed or expressed genes as unchanged. The ROC curves 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

$$\text{False positive rate} = \frac{\text{Number of false positives}}{\text{Number of Equally Expressed genes}}.$$

This is the same as the probability of Type I error denoted by $\alpha$. The false negative rate is the proportion of Differentially Expressed genes that were erroneously reported as Equally Expressed. More specifically,

$$\text{False Negative Rate} = \frac{\text{Number of false negatives}}{\text{Number of Differentially expressed genes}}.$$

This is the same as the probability of type II error.

A method whose ROC curve lies below another one is preferred (10),(3) as the curve represents the 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. The ROC curves of the methods we discussed are given in Figures 1 to 3. It can be observed that the ROC curve of the proposed method using Johnson distribution lies below the ROC curves of the other methods showing that the proposed method is better than the other methods. Better performance can also be observed as the sample size increases. Figure 3 (right panel) shows, how the ROC curves behave with the changes in shrinkage parameter,$a_0$ defined in equation 2.1. As can be seen from this figure ROC curve is better when $a_0$ is a non-zero value. Also, ROC curves behave equally good when $a_0$ is first quartile, median, third quartile or maximum, To be in safer side we took median, which is free from outliers, as the shrinkage parameter in our method.

## 3 Conclusion

A new approach is proposed for the selection of differentially expressed genes based on Johnson's system of distributions. No specific parametric form is assumed for the distribution of the expression levels. The empirical distribution is used as the common null distribution of the test statistics. The Johnson's system of distributions is used to estimate the null distribution. The simulation study shows better performance of the proposed method as compared to the conventional gene selection methods.

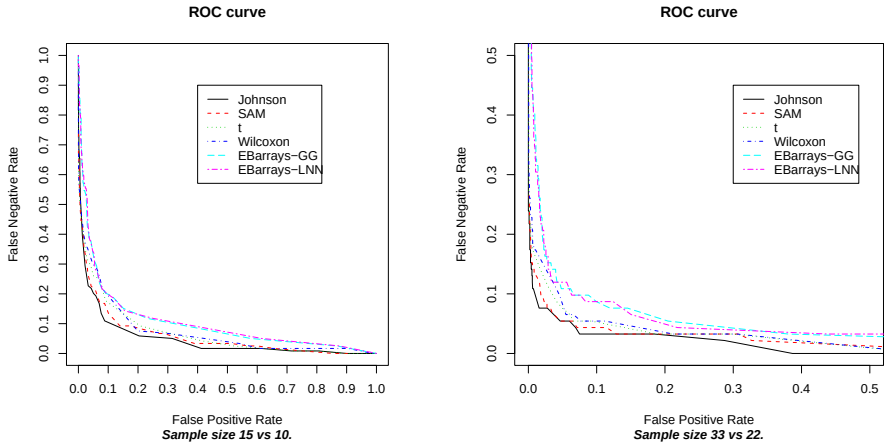


Figure 1: Receiver Operating Characteristic curve; Sample sizes - 15 vs 10 (left panel) and 33 vs 22 (right panel)

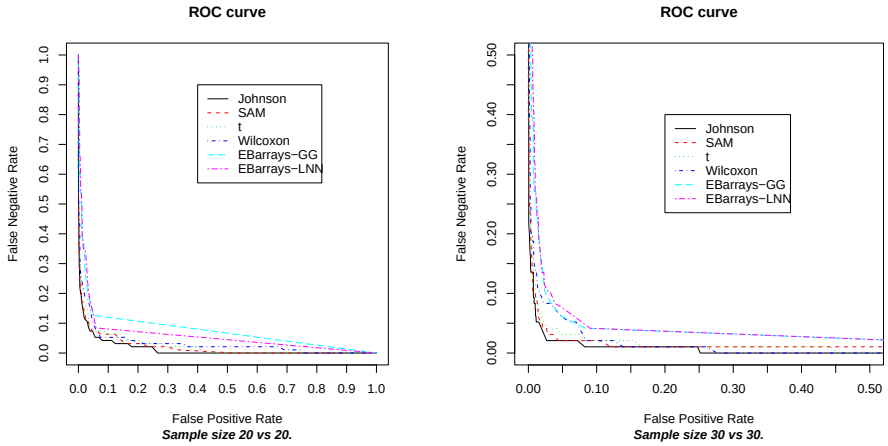


Figure 2: Receiver Operating Characteristic curve; Sample size - 20 vs 20 (left panel) and 30 vs 30 (right panel)

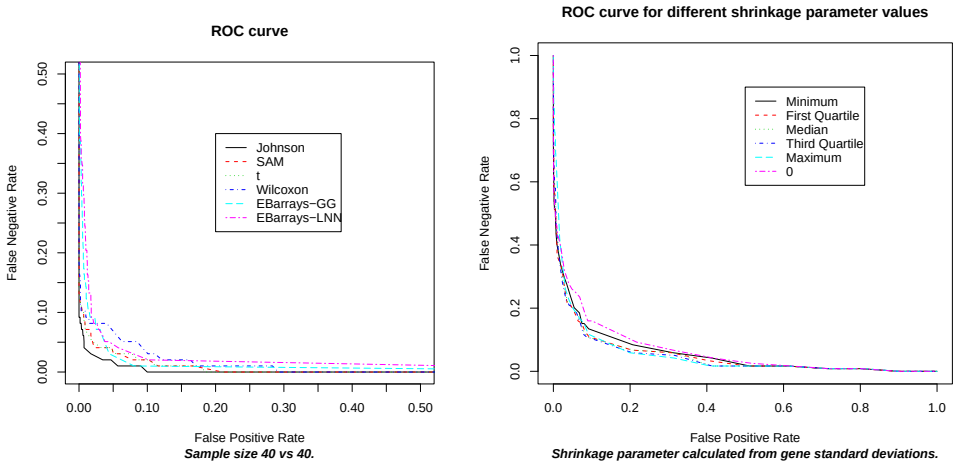


Figure 3: Receiver Operating Characteristic curve; Sample size - 40 vs 40 (left panel) and 15 vs 10 (right panel)

# 4 Acknowledgements

We acknowledge help by late Dr. A. N. V. Rao, University of South Florida in the preliminary version of this work. We would like to thank the referee for his valuable suggestions to improve this paper.

# References

[1] Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509-519.

[2] Devore, J. and Peck, R. (1997). *Statistics: the Exploration and Analysis of Data, 3rd edition.* Duxbury Press.

[3] George, F. and Ramachandran, K. M. (2007). A mixture model approach for gene selection using Johnson's system and Baye's formula. *Neural, Parallel & Scientific Computations* , **16** 45-58.

[4] Hahn, J. G., and Shapiro, S. (1967). *Statistical models in Engineering*, John Wiley and Sons.

[5] Ihaka, R., Gentleman, R. (1996). A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.

[6] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barely, Y.D., Antonellis K.J., Scherf, U. *et al* (2003). Exploration, Normlization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, **4**(2), 249-264.

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

[8] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, volume 1-2.

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

[10] Lonnstedt, I. and Speed, T. (2002). Replicated Microarray data. *Statistica Sinica*, **12**, 31-46.

[11] Markus, N. and Fred, L. (2004). Nonparametric Approaches to Detecting differentially expressed genes in replicated microarray experiments. *2nd Asia-Pacific Bioinfomatics Conference (APBC2004)*, Dunedin, New Zealand.

[12] Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R. and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of computational Biology*, **8**, 37-52.

[13] Pan, W. (2002). A Comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546-554.

[14] Slifker, J. and Shapiro, S. (1980). The Johnson System : selection and parameter estimation. *Technometrics*, **22**, 239-247.

[15] Tusher, V., Tibshirani, R., and Chu, C. (2001). Significance Analysis of microarrays applied to transcriptional response to ionizing radiation. *Proeedings of the National Academy of Sciences*, **98**, 5116-5121.

[16] Wheeler, R. (1980). Quantile Estimators of Johnson curve Parameters. *Biometrika*, **67**(3), 725-728.