Neural, Parallel, and Scientific Computations 18 (2010) 371-384

APPLICATIONS OF PENALIZED MIXTURE DISTRIBUTIONS TO MICROARRAY DATA ANALYSIS

O'NEIL LYNCH[1] , KANDETHODY M RAMACHANDRAN[2], WONKUK KIM[2]
1Mathematics Department, Minnesota State University Moorhead, Moorhead, MN 56563 ,
2 Department of Math. and Statistics, University of South Florida, Tampa, FL 33620-5700.

ABSTRACT: The main goal in analyzing microarray data is to determine the genes that are differentially expressed across two types of tissue samples or samples obtained under two experimental conditions. In this paper we propose a penalized normal mixture model (PMMM) to estimate the parameters within the framework of maximum likelihood. We penalized both the variance and the mixing proportion. The variance was penalized so that the log-likelihood will be bounded, while the mixing proportion was penalized so that we can apply the modified likelihood ratio to test for the number of components. Additionally, a weight function was introduced because the estimation method is sensitive to the presence of statistical outliers. Simulation study is conducted to demonstrate effectiveness of PMMM. Finally, the penalized method is applied to the rat data for genes in middle ear mucosa of rats with and without subacute pneumococcal middle ear infection.

1. INTRODUCTION

In recent years microarray technology has made it possible to simultaneously analyze thousands of genes. Although an enormous volume of data is being produced by microarray technologies [12, 18], one of the continuing challenges is how to analyze and interpret the large amounts of data. The methods used for such analysis, including the method of identifying genes with fold changes are known to be unreliable because in such methods the statistical variability of the data is not properly addressed [4]. While various parametric methods and tests such as the two-sample *t*-test have been applied for microarray data analysis, strong parametric assumptions made in these methods as well as strong dependency on large sample sets restrict the reliability of such techniques in microarray problems. The nonparametric statistical methods, including the Empirical Bayes (EB) method [6], the significance analysis for microarray data (SAM), [21], and mixture model method (MMM) [7, 8, 13, 16] have been applied to microarray data analysis. It is claimed that the new extensions of the (MMM) are among the available methods producing biologically-meaningful results [16]. The major disadvantages of the (MMM), is that the maximum likelihood estimates of the proportion may approach the boundary point of the parameter space and the log-likelihood approaches ∞ as the component variance approaches 0.

In this work we extend the (MMM) by penalizing the mixing proportions, the component variances and implementing a weight function. The mixing proportion was penalized so that the modified likelihood ratio tests of [2, 3] for testing the number of components of the fitted normal mixture model can be applied. The variance was penalized so that log-likelihood is bounded resulting in the existence of the MLE's. Statistical outlier distort the estimation of the parameters, therefore a weight function which gives full weight to sample points in a neighborhood of each component mean, but automatically reduced weights to sample points not in that neighborhood was implemented.

This paper is organized as follows. Section 2 describes the methodology, gives the model fitting algorithm. In section 3, we explain the proposed modifications. In section 4, the proposed method is applied to the rat data of [16], containing expression levels of 1176 genes of rats with and without pneumococcal middle ear infection, in addition to doing a simulation study. The results are compared to that of SAM. In section 5 we present concluding remarks.

Received April 25, 2010

2 MMM METHOD

We start this section with a brief review of the existing mixture method techniques [16]. Let Y_{ik} be the expression level of gene *i* in array *k* ($i=1,...,N;\ k=1,...,K_1, K_1+1,...,K_1+K_2$). Suppose that the first K_1 and the last K_2 arrays are both obtained under two different conditions. A linear statistical model is ([16])

$$Y_{ik} = a_i + b_i x_k + \varepsilon_{ik} \tag{1}$$

where $x_k = 1$ for $1 \le k \le K_1$ and $x_k = 0$ for $K_1 + 1 \le k \le K_1 + K_2$, and ε_{ik} are random errors with mean 0. We do not assume the homoscedastic variances. Hence $a_i + b_i$ and a_i are the mean expression levels of gene *i* under the two conditions, respectively. For simplicity of the analysis we assume that both K_1 and K_2 the number of replications for each experimental condition is even. Testing difference in the mean expression levels under the two conditions is equivalent to testing for the null hypothesis

$$H_0 : b_i = 0 \text{ versus } H_1 : b_i \neq 0 \tag{2}$$

The following *t*-statistic type scores z_i and Z_i are calculated from the data [16]. Define a_i to be a column vector containing random permutation of $K_1/2$: 1's and b_i to be a column vector containing random permutation of $K_2/2$: -1's. Let

$$z_i = \frac{Y_{i(1)} a_i / K_1 + Y_{i(2)} b_i / K_2}{\sqrt{v_{(1),i}/K_1 + v_{(2),i}/K_2}} \sim f_0 \tag{3}$$

$$Z_i = \frac{\sum_{k=1}^{K_1} Y_{ik}/K_1 - \sum_{k=K_1+1}^{K_1+K_2} Y_{ik}/K_2}{\sqrt{v_{(1),i}/K_1 + v_{(2),i}/K_2}} = \frac{\overline{Y}_{(1),i} - \overline{Y}_{(2),i}}{\sqrt{v_{(1),i}/K_1 + v_{(2),i}/K_2}} \sim f_1 \tag{4}$$

where $v_{(1),i} = (K_1 - 1)^{-1} \sum_{k=1}^{K_1} (Y_{ik} - \overline{Y}_{(1),i})^2$ and $v_{(2),i} = (K_2 - 1)^{-1} \sum_{k=K_1+1}^{K_2} (Y_{ik} - \overline{Y}_{(2),i})^2$ are the sample variances. Since z_i and Z_i are not assumed to be normally distributed, the distribution function of f_0 and f_1 are estimated using mixture of normal distributions as opposed to using kernel density estimation. Therefore, f_0 and f_1 are estimated as:

$$f_0(z; \Psi_{g_0}) = \sum_{j=1}^{g_0} \pi_j \phi(z; \mu_j, \sigma_j^2) \tag{5}$$

$$f_1(Z; \Psi_g) = \sum_{j=1}^{g} \pi_j \phi(Z; \mu_j, \sigma_j^2) \tag{6}$$

where $\phi(.; \mu_j, \sigma_j^2)$ denotes the normal density function with mean $-\infty < \mu_j < \infty$, variance $\sigma_j^2 > 0$ and $0 \le \pi_j \le 1$ denotes the mixing proportion such that $\sum_{j=1}^{p} \pi_j = 1$ (where $p = g_0$ or g). Additionally, $\Psi = \Psi_p$ represents all unknown parameters $\{(\pi_j, \mu_j, \sigma_j^2) : j = 1,..., p\}$ in a *p*-component mixture model. Next, we describe how to fit the normal mixture model.

Mixture model is usually fitted by maximum likelihood estimations using the expectation-maximization (EM) algorithm [5]. Given *N* observations $z_1,..., z_N$, maximize the log-likelihood

$$\log L(\Psi_{g_0}) = \sum_{i=1}^{N} \log f_0(z_i; \Psi_{g_0}) \qquad (7)$$

to obtain the maximum likelihood estimate Ψ_{g_0} for the distribution f_0. The EM algorithm can be used to compute Ψ_{g_0} iteratively through the following steps.

E-Step

$$\pi_{ij}(\Psi^{\{t+1\}}) = \left. \frac{\pi_j \phi(z_i; \mu_j, \sigma_j^2)}{\sum_{j=1}^{g_0} \pi_j \phi(z_i; \mu_j, \sigma_j^2)} \right|_{\Psi^{\{t\}}} \qquad (8)$$

M-Step

$$\pi_j^{\{t+1\}} = \frac{\sum_{i=1}^{N} \pi_{ij}(\Psi^{\{t\}})}{N} \qquad (9)$$

$$\mu_j^{\{t+1\}} = \frac{\sum_{i=1}^{N} \pi_{ij} z_i}{\sum_{i=1}^{N} \pi_{ij}} \qquad (10)$$

$$(\sigma_j^2)^{\{t+1\}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{g_0} \pi_{ij} \left(z_i - \frac{\sum_{i=1}^{N} \pi_{ij} z_i}{\sum_{i=1}^{N} \pi_{ij}}\right)^2}{N} \qquad (11)$$

Equation (8) the E-Step is the posterior probability that z_i belongs to the *j*th-component of the mixture. At convergence, we obtain the maximum likelihood estimate of Ψ_{g_0}. After finding the optimized Ψ_{g_0} for different g_0's the algorithm selects the sub-optimal g_0 corresponding to the first local minimum of the AIC or BIC [9],

$$AIC = -2 \log L(\Psi_{g_0}) + 2v_{g_0},$$

$$BIC = -2 \log L(\Psi_{g_0}) + v_{g_0} \log(N),$$

where v_{g_0} is the number of independent parameters in Φ_{g_0}. Then the algorithm uses the resulting g_0 as the number of normal functions to fit f_0. The same procedure can be applied to estimate f_1. As mentioned above, with the fixed number of normal functions, the parameters of f_0 and f_1 are iteratively updated for a number of iterations. When the iterations are terminated, the likelihood ratio of (12) is estimated based on the final estimations of f_0 and f_1.

One of the problems with this method is that the AIC and BIC may not agree with each other in some cases, therefore it often means that several models are reasonable and that no one can dominate the others. Therefore we seek other methods which are more reliable in the selection of g_0, the number of components, as in [2,3].

In order to determine the statistical significance, we want to test for the null hypothesis H_0 that *Z* is from f_0. Construct a likelihood ratio test (LRT) based on the following statistic,

$$LR(Z) = [f_0(Z) / f_1(Z)]. \qquad (12)$$

A large value of *LR*(*Z*) gives no evidence against H_0, whereas a too small value of *LR*(*Z*) leads to rejecting H_0. With the normal mixture model, it is possible to numerically determine the rejection region. For any given false positive rate α (In the literature of microarray processing,

$\alpha = 0.01$ is often used as the genome wide significant level), we can use the bisection method [17] to solve $\alpha = \int_{LR(z)<s} f_0(z)dz$ and obtain the suitable cut off point $s = s(\alpha)$. Then the rejection region is $RR(\alpha) = \{Z : LR(Z) < s\}$. This method of using the LRT in MMM is called as MMM-LRT, [16]. Similar to SAM [21], we can estimate the numbers of false positive (*FP*) and total positive (*TP*) directly. In MMM-LRT, for any given *s*, we have:

$$FP(s) = \frac{1}{B}\sum_{b=1}^{B} n(i : LR(z_i^{(b)}) < s), \quad TP(s) = n(i : LR(Z_i) < s)$$

Here $n(i)$ represents the number of genes. Based on the estimated *FP* and *TP*, we can also calculate the false discovery rate as *FDR* = *FP/TP* ([1, 7,16, 21]). The existing approach used the AIC and/or BIC as a criteria for model selection but for the model selection we used the modified likelihood ratio test to test the hypotheses: a 2-component model (alternative hypothesis) vs. 1-component (null hypothesis) and 3-component model (alternative hypothesis) vs. 2-component (null hypothesis), [2, 3]. Define

$$l_N(\Psi_{g_0} | z) = \sum_{i=1}^{N} \ln\left\{\sum_{j=1}^{g_0} \pi_j f_{ij}(z_i | \theta)\right\} + C\sum_{j=1}^{g_0} \ln(g_0 \pi_j).$$

Below we present the theorems of Chen et al. [2, 3] to carry out these tests of hypothesis. Theorem 1 is the test for a 2-component model vs. 1-component, while Theorem 2 is the test for a 3-component model vs. 2-component.

Theorem 1 *([2]) If the regularity conditions hold, the asymptotic null distribution of the modified LRT statistics*

$M_n = 2\{l_n(\pi, \theta_1, \theta_2) - l_n(1/2, \theta, \theta)\}$ *for testing a 2-component (alternative) vs. 1-component (null), is the mixture of* χ_1^2 *and* χ_0^2 *with equal weights, i.e.* $(1/2)\chi_0^2 + (1/2)\chi_1^2$, *where* χ_0^2 *is a degenerate distribution with all its mass at 0.*

Theorem 2 *([3]) If the regularity conditions hold, and the true distribution is a 2-component model. Then the asymptotic null distribution of the modified LRT statistics* $R_n = 2\{l_n(\pi_1, \pi_2, \theta_1, \theta_2, \theta_3) - l_n(\pi, \theta_1, \theta_2)\}$ *for testing a 3-component (alternative) vs. 2-component (null), is the mixture of* $((1/2) - (\alpha / 2\pi))\chi_0^2 + (1/2)\chi_1^2 + (\alpha / 2\pi)\chi_2^2$, *where* $\alpha = \cos^{-1}(\rho)$, ρ *is the correlation coefficient between the two components of the null hypothesis and* χ_0^2 *is a degenerate distribution with all its mass at 0.*

Additionally since the model is sensitive to the presence of outlier, Tadjudin et al. [20] discussed how a weight function ω_{ij} given by $\omega_{ij} = 1$ for $0 \le d_{ij} \le 3$, and $\omega_{ij} = 3 / d_{ij}$ for $3 < d_{ij} < \infty$, where $d_{ij} = (y_i - \mu_j) / \sigma_j$ could assign each observation a measure of typicality for each component.

3. PENALIZATION OF MMM MODEL

We saw that the mixture model used unequal variances for each component. Keifer and Wolfowitz [11] showed that when applying mixture of normals with unequal variances in each component the likelihood approaches ∞ as one of the variances approaches 0. In order to see how well a homoscedastic mixture model fit a heteroscedastic data, we simulated mixture distributions from a sample of size N = 500 from

$$Z \sim 0.5\phi(y|4,1) + 0.5\phi(y|8,2)$$

And fitted the simulated data with equal and unequal variances. The result is given in the figure 1, where the dotted and bold lines represent the heteroscedastic and homoscedastic models respectively.

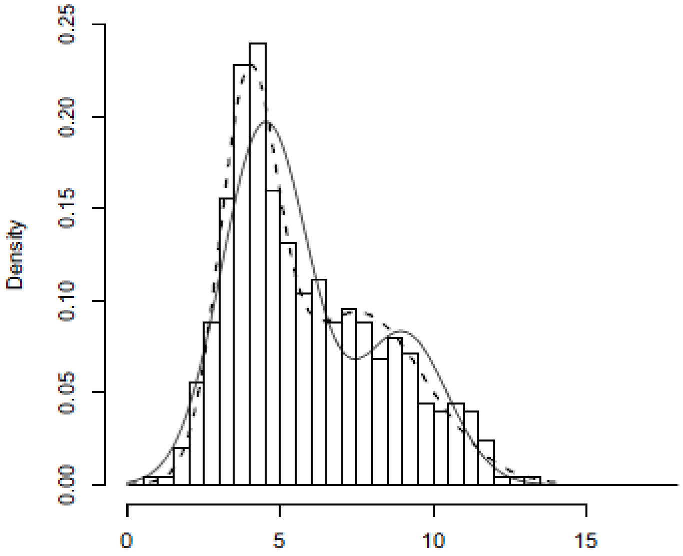

Figure 1: Histogram, heteroscedastic and homoscedastic fit for simulated data from the mixture $0.5\phi(y|4,1)+0.5\phi(y|8,2)$

Another 500 data is generated from and fitted and we got the result as:

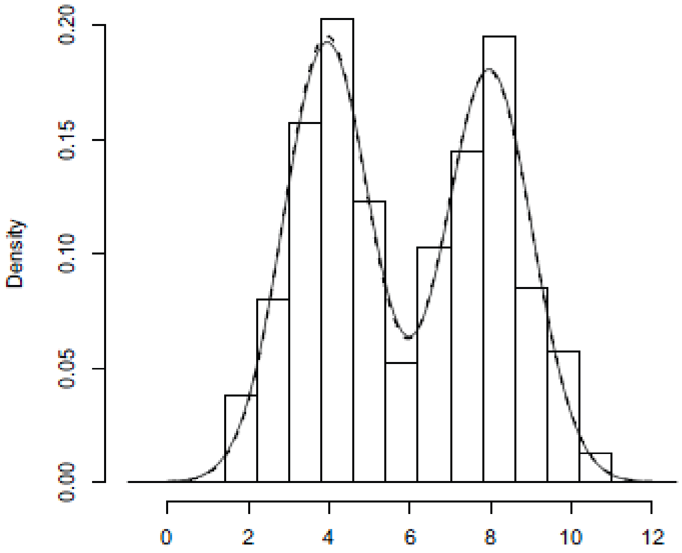

Figure 2: Histogram, heteroscedastic and homoscedastic fit for simulated data from the mixture $0.5\phi(y|4,1)+0.5\phi(y|8,1)$

The model with unequal variances seems to be a better fit in the case where the simulated data with unequal variance was fitted with unequal variance method as oppose to when fitted using equal variance based method(Figure 2). However, there was no difference if we fitted normal mixture model with equal or unequal variances for the data that was simulated using equal variance (Figure 1).

Another problem is that the maximum likelihood estimates of the proportions π_j can be very close to the boundary point 0.

To overcome these problems and to determine the distributions of f_0 and f_1 we maximize the penalized log-likelihood:

$$\log L(\Psi_p) = \sum_{i=1}^{N} \log f(z_i; \Psi_p) + C\sum_{j=1}^{p} \log(p\pi_j) + \sum_{j=1}^{p} \log h(\sigma_j) \tag{13}$$

(where $f = f_0$ or f_1) to obtain the maximum likelihood estimates Φ_p ($p = g_0$ or g) of the unknown parameters $\left\{(\pi_j, \mu_j, \sigma_j^2): j = 1, ..., p\right\}$, where $C > 0$ is a constant. The second and third terms on the right hand side of equation (13) are the penalty terms for the proportion (so that the modified likelihood test can be applied) and variance respectively. From [2] it was mentioned that an appropriate choice for C is $C = \log(M)$, when the parameter μ in the kernel density is restricted to $[-M, M]$. Furthermore, from simulation, we observe that the method is not sensitive to the value of C and the choice of $C = log(M)$ works well.
We will choose $h(.)$ such that

(i) $\lim_{\sigma \to 0} \frac{1}{\sigma^N} h(\sigma) = 0$, *for all* N *so that the penalized MLE exists.*

In order to prove the consistency it is required that h also satisfied the following conditions:
(ii) $h(\sigma)$ *is many-to-one from* $(0,1)$ *onto* $(0, G)$, $G > 0$,
(iii) h is strictly increasing in an open interval $(0, \delta)$ *of the origin which has a non-null measure,*
(iv) h is continuously differentiable on $(0, \infty)$.

One such distribution that satisfy the aforementioned conditions on $h(\sigma_j)$ is the inverse gamma function, $h(\sigma_j) = \frac{\alpha^\beta}{\Gamma(\beta)} \frac{1}{\sigma_j^{2(\beta+1)}} \exp\{-\frac{\alpha}{\sigma_j^2}\}$, $\alpha > 0$, $\beta > 0$. Hence for the further analysis, we assume that $h(\sigma)$ has inverse gamma distribution. It should be noted that similar analysis can be done for inverse chi-square penalty function, [14].

For $p = g_0$ or g component mixture model, parameters can be estimated using similar steps to (7) with (8), (9) and (10) modified, the EM algorithm computes Ψ_p by iterating steps [14]: for $j = 1, ..., p$,

$$\pi_j^{\{t+1\}} = \frac{\sum_{i=1}^{N} \pi_{ij}^{\{t\}} + C}{N + pC},$$

$$\mu_j^{\{t+1\}} = \frac{\sum_{i=1}^{N} \pi_{ij}^{\{t\}} \omega_{ij}^{t} z_i}{\sum_{i=1}^{N} \pi_{ij}^{\{t\}}}$$

$$(\sigma_j^2)^{\{t+1\}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{p}\pi_{ij}^{\{t\}}(\omega_{ij}^2)^t(z_i - \frac{\sum_{i=1}^{N}\pi_{ij}^{\{t\}}z_i}{\sum_{i=1}^{N}\pi_{ij}^{\{t\}}})^2 + 2\alpha}{N + 2(\beta+1)} \quad ,$$

and

$$\omega_{ij}^{\{t+1\}} = \frac{3\sigma_j^t}{(y_i - \mu_j^t)} \; .$$

Consistency and Asymptotic Normality

Let $Y_1,...,Y_N$ be a random sample of size N from the mixture model with density given by (5), where the parameters $\psi \in \Psi$ and let $\overline{\Psi}$ denote the closure of set Ψ. The likelihood function is unbounded on Ψ, [14]. This was circumvented by adding a penalty term for the variance parameter. Let $H = L^1(\phi(y,\psi_0))$ with norm $\|\phi\| = \int \phi$. Under this norm, H is a Banach space. Let E_H denotes the expectation in the space H. Consider L_N that is the extension of L to $\overline{\Psi}$, i.e.,

$$L_N = \begin{cases} 0 & \text{if } \sigma_k = 0, \infty \text{ or } \mu_\text{k} = \pm\infty \\ f(y_1,...,y_N|\psi)\prod_{j=1}^{g_0} h(\sigma_j)\prod_{j=1}^{g_0}(g_0\pi_j)^C & \text{otherwise} \end{cases}$$

Now, we state the strong consistency of the penalized MLE by means of the following two Theorems. We will refer to [14] for the complete proofs of these results.

Theorem 3: *Let S be a closed subset of $\overline{\Psi}$ such that*

$$S = \{\psi \in \overline{\Psi} | \exists\{1,..g_0\} \text{ so that } \sigma_\text{j} \in [0,\eta)\}$$

and such that $\psi_0 \notin S$. Then

$$P(\lim_{N\to\infty}\sup_{\psi\in S}\frac{L_N(Y_1,...,Y_N|\psi)}{L_N(Y_1,...,Y_N|\psi_0)} = 0) = 1\,.$$

Theorem 4. *Let $\overline{\psi}_N = \overline{\psi}(Y_1,...,Y_N) \in \overline{\Psi}$ be a function of $Y_1,...,Y_N$ such that*

$$\frac{L_N(Y_1,...,Y_N|\overline{\psi}_N)}{L_N(Y_1,...,Y_N|\psi_0)} \geq \rho > 0, \forall Y_1,...,Y_N,\ \forall N$$

Then

$$P(\lim_{N\to\infty}\overline{\psi}_N = \psi_0) = 1\,.$$

From the previous result, by considering $\rho = 1$, we obtain the following corollary.

Corollary 5. *The penalized maximum likelihood estimator is strongly consistent, i.e. the point $\overline{\psi}_N$ which maximizes L_N is such that $\overline{\psi}_N \to \psi_0$ a.s.*

For the speed of convergence of the penalized estimator, we have following result.

Theorem 6. *Assume that the parameters satisfy following condition*

$$(\mu_k, \sigma_k) \neq (\mu_m, \sigma_m) \text{ for } k \neq m, \forall k = 1, ..., g$$

and the penalizing function is such that $\dfrac{h^s(\sigma)}{h(\sigma)}$ *is bounded for* $s = 1, 2, 3$ and $\forall \sigma \in \{\sigma_{01}, ..., \sigma_{0N}\}$ *then* $\sqrt{N}(\overline{\psi}_N - \psi_0)$ *is asymptotically normal distributed with mean zero and covariance matrix* $I(\psi_0)^{-1}$*, where the information matrix*

$$I(\psi_0) = E_H[(\frac{\partial \ln \phi(\psi_0)}{\partial \psi})(\frac{\partial \ln \phi(\psi_0)}{\partial \psi})^T].$$

The distribution of f is estimated exactly as was done in the previous section and we then compared f_0 and f_1 by likelihood ratio discussed in section 2. However, since the regularity conditions necessary for Theorems 1 and 2 do not satisfy for the PMMM model (13), therefore the asymptotic distribution is not chi-squared with degrees of freedom 2. The theoretical distribution of the penalized modified likelihood ratio statistic in the case of unequal variances for each component is an open problem. Instead we introduce the following simulation method based on a regression model as a function of $(1/N)^t$ given by

$$E(z_{PNs}) = \alpha_{P,t} + b_{P,t}(1/N)^t,$$

Where z_{PNs} is the P^{th} percentile of the s^{th} subsample of size N. The simulation of the null distribution is done as follows. In the case of null hypothesis is that data comes from single normal distribution against the alternative that it is mixture of two heterogeneous normals, we simulated 500 replicates of the standard normal N(0, 1) of sample sizes 100, 250, 500, 750 and 1000. Then we fitted 2-components normal mixture models for each of the sample sizes and calculated the penalized modified log likelihood ratio test (PMLRT) define as

$$R_N = 2\{\ln L(\pi, \theta_1, \theta_2, \sigma_1, \sigma_2) - \ln L(1/2, \theta, \theta, \sigma, \sigma)$$

A linear model was fitted using the 5 values of PMLRT to determine the degrees of freedom, and we obtained the degrees of freedom of the simulated chi-squared null distribution as

$$f = 2.8 + 13.8N^{-0.5} \tag{14}$$

Table 1 shows the mean, variance and percentiles of the PMLRT for the sample sizes 100, 250, 500, 750 and 1000 for hypothesis. The percentiles in brackets are that of the chi-squared distribution with degrees of freedom given by (14) (which of course is a gamma distribution with mean $1.4 + 6.9N^{-1/2}$ and second parameter 0.5), while those percentiles not in brackets are the ordered simulated percentiles of
PMLRT.

Table 1: *Mean, variance and percentiles for the penalized modified likelihood, based on 500 replicates for each sample for testing the hypothesis a 1-component against 2-components.*

Sample size	100	250	500	750	1000
Mean	4.07	3.89	3.50	3.15	3.19
Variance	8.08	8.14	7.50	6.80	7.06
Percentiles					
50%	3.30(3.53)	3.12(3.03)	2.88(2.78)	2.51(2.67)	2.49(2.60)
75%	5.65(5.61)	5.24(4.97)	4.63(4.64)	4.07(4.50)	4.25(4.41)
90%	8.11(8.05)	7.95(7.29)	7.00(6.90)	6.27(6.72)	6.88(6.62)
95%	9.81(9.78)	9.31(8.95)	8.73(8.53)	7.96(8.33)	8.34(8.22)

For the testing of two components versus three components, similar simulation resulted in the linear regression model for the degrees of freedom as a function of N being

$$f = 4.9 + 13.4N^{-1/2}$$

And the corresponding results are given in Table 2.

Table 2: *Mean, variance and percentiles for the penalized modified likelihood, based on 500 replicates for each sample for testing the hypothesis 2-components against 3-components.*

Sample size	100	250	500	750	1000
Mean	6.13	5.93	5.63	5.28	5.21
Variance	12.03	12.13	11.85	10.84	10.63
Percentiles					
50%	5.41(5.59)	5.19(5.10)	4.95(4.85)	4.66(4.74)	4.59(4.67)
75%	8.07(8.13)	7.85(7.53)	7.38(7.23)	7.25(7.10)	7.19(7.02)
90%	11.25(10.98)	10.73(10.29)	10.11(9.94)	9.82(9.79)	9.33(9.70)
95%	12.88(12.95)	12.35(12.21)	11.97(11.84)	11.85(11.67)	11.72(11.57)

4. DATA APPLICATION

Simulation

To mimic the real gene data, we generated data for N = 1176 genes under the following setup. We assume that $K_1 = 2$*;* $K_2 = 6$ and there are 200 differentially expressed (DE) genes. The data for the equally expressed (EE) genes are simulated from $N(\mu_{i1}, \sigma_{i1}^2)$ for $i = 1, ..., K_1$ and $N(\mu_{i2}, \sigma_{i2}^2)$ for $i = K_1 + 1, ..., K_1 + K_2$, where $\mu_{i1} = \mu_{i1} \sim N(0, 2)$ and σ_{i1} and σ_{i2} are generated from *Gamma*(2*;* 4), respectively. Note that such generated σ_{i1} and σ_{i2} take different values for each gene and are also different between genes. The data for DE genes were generated similarly. However, in this case, μ_{i1} and μ_{i2} were generated from *N*(0*;* 2) separately. The standard deviations σ_{i1} and σ_{i2} are generated the same way as in the EE gene case.

In our method we fitted a 1, 2 and 3-component normal mixture model and calculated M_n and R_n as defined in Theorems 1 and 2 respectively, modified to $L_N(\Psi_p)$ defined through (13).

Table 3: *MRLT for fitted normal mixture models of z and Z*

	M_n	R_n
f_0	*4.56 (P<0.01)*	*1.22 (P>0.05)*
f	*6.78 (P<0.01)*	*1.06 (P>0.05)*

Table 3 displays the results and we therefore choose the 2-component normal mixture model for both f_0 and *f* which are stated below:

$$f_0(z) = 0.01\phi(z; -0.287, 0.05673^2) + 0.99\phi(z; -0.004558, 0.40812^2)$$

$$f(z) = 0.20\phi(z; -2.442, 0.43703^2) + 0.80\phi(z; -0.0062961, 0.48583^2).$$

Figure 3(a) shows the histograms of z with the fitted normal mixture models, which shows strong agreement. Similar observation for Z is shown in 3(b) with the dotted line being that of the fitted mixture model of f_0. Figure 4 illustrates the likelihood ratio statistic as a function of the Z values.

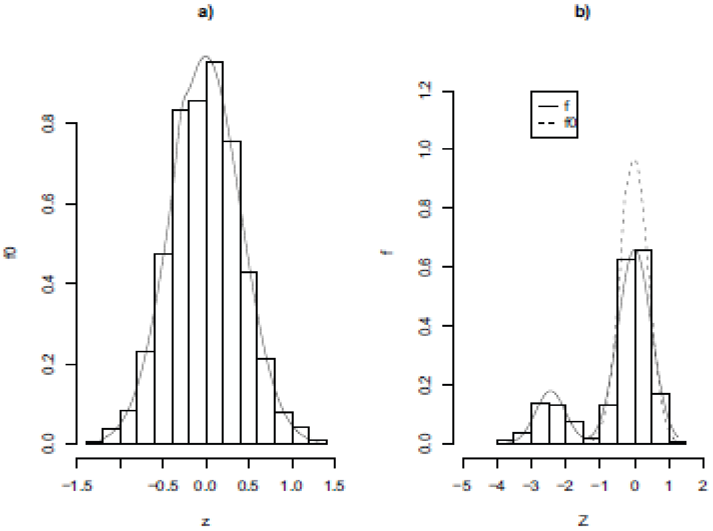

Figure 3: *Histogram of* z *and* Z *and fitted models for the simulated data*

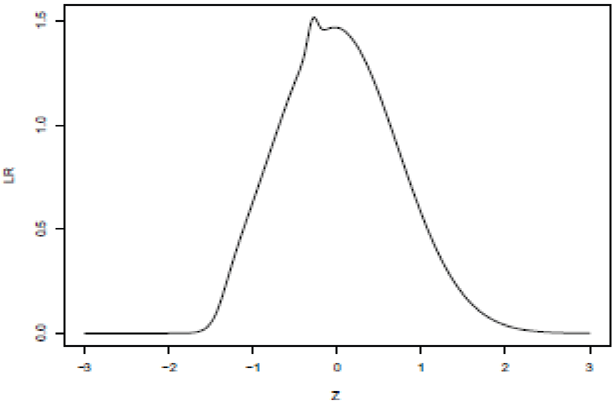

Figure 4: *The likelihood ratio statistic as a function of* Z *value for the simulated data*

Table 4: *Values of TP, FP and FDR from PMMM*

s	*MedianFP*	*MeanFP*	*TP*	*FDR%*
0.07	0	0.069	196	0.00
0.10	0	0.138	196	0.00
0.15	0	0.310	196	0.00
0.30	2	1.655	201	1.00
0.35	3	3.828	203	1.48
0.40	5	5.621	205	2.44
0.45	14	13.931	210	6.67
0.60	26	25.724	221	11.76
0.70	43	43.207	231	18.61
0.90	68	66.966	248	27.42
1.00	104	103.517	270	38.52

For our method the median number of *FP* were calculated from the null scores of B = 29 permutations of the data set. For SAM, all the results were obtained from the R-package sam3.0. For the purpose of comparison,

Table 5: *Values of TP, FP and FDR from SAM*

Δ	*Median FP*	*Mean FP*	*TP*	*FDR%*
0.49	3.71	6.496	195	1.90
0.47	4.64	6.496	197	2.36
0.45	6.50	8.352	198	3.28
0.43	7.42	10.208	200	3.71
0.42	8.35	12.064	203	4.11
0.37	11.14	14.848	206	5.41
0.32	23.20	27.840	211	11.00
0.28	33.41	40.832	221	15.12
0.25	45.47	55.680	230	19.77
0.20	69.60	82.592	246	28.29
0.16	107.65	118.042	268	40.17

the cut-off points *s* used in our method are specifically chosen to match the number of *TP* produced by sam3.0. It is seen from Tables 4 and 5 that our method outperform SAM. Figure 5 displayed a graphical comparison of the numerical results presented in Tables 4 and 5.

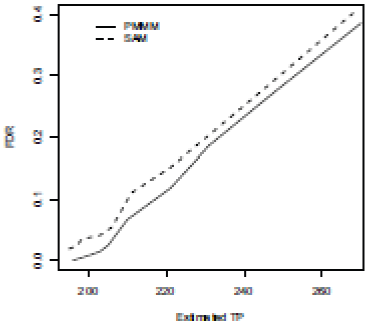

Figure 5: *The values of FDR from PMMM method and SAM for the simulated data*

Real Data Example

In this section, we apply the penalized modified likelihood method to the rat data of [15]. The data is from a study, that applied radioactively labeled DNA microarrays ([10]) to the mRNA analysis of 1,176 genes in middle ear mucosa of rats with and without subacute pneumococcal middle ear infection. The data consists of eight experiments: two DNA microarrays were run with controls while six were run with pneumococcal middle ear infection. The data was processed by first taking a natural logarithm transformation for all the observed gene expression levels so that the resulting data is less skewed. Then, for each microarray, we standardize the transformed gene expression levels by subtracting their mean and dividing by their standard deviation.

Table 6: *MRLT for fitted normal mixture models of z and Z*

	M_n	R_n
f_0	3.19 (P < 0:01)	0.92 (P > 0:05)
f	3.54 (P < 0:01)	1.16 (P > 0:05)

From Table 6 we choose the 2-component normal mixture model for both f_0 and f which are stated below:

$$f_0(z) = 0.983\phi(z; 0.011, 0.185) + 0.017\phi(z; 0.297, 0.069)$$

$$f(z) = 0.958\phi(z; -0.032, 0.539) + 0.042\phi(z; 0.246, 0.004).$$

Table 7: *Values of TP, FP and FDR from PMMM*

s	*MedianFP*	*MeanFP*	*TP*	*FDR%*
0.07	0	0.03	94	0.00
0.10	0	0.07	103	0.00
0.15	0	0.28	113	0.00
0.30	3	3.17	144	2.08
0.35	8	8.75	168	4.76
0.40	12	12.44	178	6.74
0.45	29	27.86	215	13.49
0.60	44	46.17	248	17.74
0.70	65	65.96	288	22.57
0.90	95	95.59	323	29.41
1.00	134	133.83	368	36.41

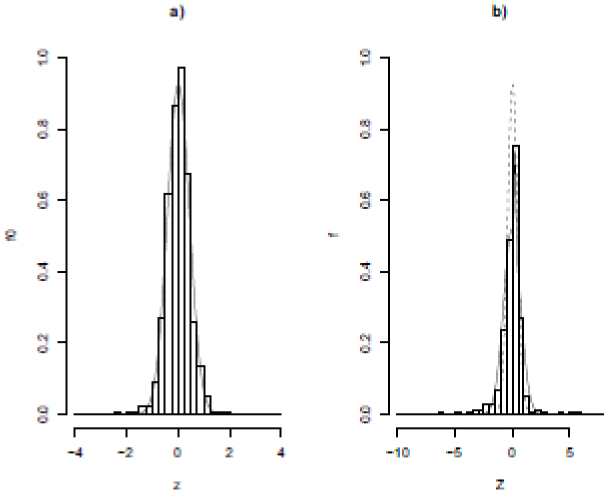

Figure 6: Histograms of z , Z , and fitted models for the rat data

Figure 6(a) presented the histogram of z and the fitted f_0, which do not indicate strong discrepancy. The histogram of Z and the fitted mixture model are shown in Figure 6(b) with f_0 shown in the dotted line. The constructed LR statistics are plotted in Figure 7. It is not surprising to see as Z moves away from *0*, *LR(Z)* decreases.

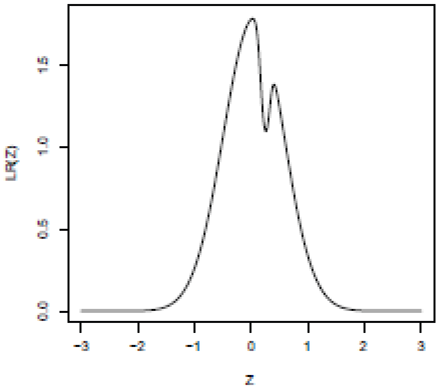

Figure 7: *The likelihood ratio curve for the rat data*

Table 8: *Values of TP, FP and FDR from SAM*

Δ	*Median FP*	*Mean FP*	*TP*	*FDR%*
0.94	4.2	16.73	80	5.23
0.88	9.1	23.71	101	8.97
0.78	11.2	29.98	149	9.96
0.68	19.5	45.32	149	13.10
0.63	25.1	62.76	168	14.94
0.58	34.2	76.70	198	17.26
0.54	49.5	97.62	238	20.80
0.50	57.2	109.47	259	22.08
0.46	75.7	135.27	301	25.13
0.42	93.1	167.35	336	27.70
0.38	132.5	221.04	420	31.54

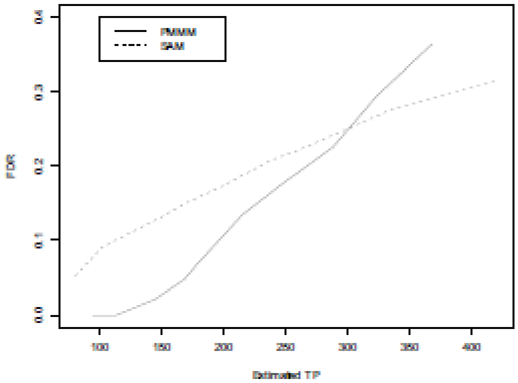

Figure 8: *The values of FDR from PMM method and SAM for the rat data*

Tables 7 and 8 report the results from our method and SAM. Figure 8 displays the *FDR* with respect to different values of *TP*. For $TP \leq 300$, the advantage of our method over SAM is obvious. For *TP* > 300, the *FDR* value of our method is higher than that of SAM. It is noteworthy that for this data set the number of genes that one wants to detect should be no greater than 300, hence the PMMM approach provides statistical significant results compared to that of SAM.

5. CONCLUSION

In this paper we have proposed an improved method of determining the number of components of the distributions of the two *t*-statistic-type scores by applying the modified likelihood ratio test of [2,3]. We also implemented a penalty term for the variance so that the log-likelihood will be bounded. We demonstrated that the testing procedure using our method has a higher power (or lower FDR) than that from the most popularly used method, SAM. Further details are available in [14].

REFERENCES

[1] Benjamini, Y., and Hochberg, Y., (1995), *Controlling the false discovery rate: a practical and powerful approach to multiple testing,* J. R. Statist. Soc., Vol. 57, pp. 289-300.

[2] Chen, H., Chen, J., and Kalbfleischd, J. D., (2001), *A Modified Likelihood ratio Test for Homogeneity in Finite Mixture Models,* J. R. Statist. Soc., Vol. 63, Part 1, pp. 19-29.

[3] Chen, H., Chen, J., and Kalbfleischd, J. D., (2004), *Testing for a finite mixture model with two components,* J. R. Statist. Soc., Vol. 66, Part 1, pp. 95-115.

[4] Chen, Y., Doughterty, E., and Bitter, M., (1997), *Ratio-based decisions and the quantitative analysis of cDNA microarray images*, J. Biomedical Optics, Vol. 2, pp. 364-367.

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977), *Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)*, J. R. Statist. Soc. B, Vol. 39, pp. 1-38.

[6] Efron, B., Tibshirani, R. J., Tusher, V., (2001), *Empirical Bayes Analysis of a Microarray Experiment*, J. of the Amer. Stat. Ass., Vol. 96, pp. 1151-1160.

[7] Florence George, Kandethody M Ramachandran, and Li Lihua, Gene Selection with Johnson's Distribution, in *Journal of Statistical Research*, Vol. **43**, No. 1, pp. 117-125, 2009.

[8] Florence George and Kandethody M Ramachandran, "A Mixture Model approach for Gene selection using Johnson's system and Bayes formula", in Neural, Parallel, and Scientific Computations, Vol. 16, no. 1, pp. 45-57, 2008.

[9] Fraley, C. and Raftery, A. E., (1998), *How many cluters? Which clustering methods? - Answer via model-based cluster analysis.*, The Computer Journal, Vol. 41, pp. 578-588.

[10] Friemert, C., Erfle, V., Strauss, G., 1998. Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression, Methods Mol. Cell Biol. 1, 143–153.

[11] Kiefer, J., Wolfowitz, J., (1956), *Consistency of the maximum-likelihood estimator in the presence of infinitely many incidental parameters*, Ann. Math. Stat., Vol 27, pp. 888-906.

[12] Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E. L., (1996), *Expression of monitoring by hybridization to high-density oligonucleotide arrays*, Nature Biotechnology, Vol 14, 1675- 1996.

[13] Najarian, K., Zaheri, M., Rad, A. A., Najarian, S. and Dargahi, J., (2007) *A novel Mixture Model Methiod for identification of differentially expressed genes from DNA microarray data*, Bioinformatics, Vol 5, pp. 201-211.

[14] O'Neil Lynch (2009), Mixture distributions with application to microarray data analysis, Ph.D. dissertation, University of South Florida.

[15] Pan, W., (2002), *A comparitive review of statistical methods for dicovering differentially expressed genes in replicated microarray experiments*, Bioinformatics, Vol 27, pp. 546-554.

[16] Pan, W., Lin, J., Le, C., (2003), *A Mixture Model Approach to Detecting Differentially Expressed Genes with Microarray Data*, Functional & Integrative Genomics, Vol 3, pp. 117-124.

[17] Press, W. H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B. P., (1992), *Numerical Recipes in C, The Art of Scientific Computing*, 2nd ed. New York: Cambridge University Press.

[18] Schena, M., Shalon, D., Davis, R. W. and Brown, P. O., (1995), *Quantitative monitoring of gene expression patterns with a complementary DNA microarray*, Science, Vol. 270, pp. 467-470.

[19] Schwartz, G., (1978), *Estimating the dimensions of a model*, Annals of Statistics, Vol. 6, pp. 461-464.

[20] Tadjudi, S., and Landgrebe, D. A., (2000), *Robust Parameter Estimation For Mixture Model*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 38, No. 1, pp. 439-445.

[21] Tusher, V., Tibshirani, R. J., Chu, G., (2001), *Significant Analysis of Microarrays Applied to the Ionizing Radiation Response.*, Proc. Nat. Acad. Sci., Vol. 98, pp. 5116-5121.